
KTH Computer Science
and Communication

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

HÅKON ANDREAS HOEL

Doctoral Thesis
Stockholm, Sweden, 2012

TRITA-CSC-A 2012:06
ISSN-1653-5723
ISRN KTH/CSC/A–12/06-SE
ISBN 978-91-7501-350-3

School of Computer Science and Communication
KTH
SE-100 44 Stockholm
SWEDEN

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:

**Paper I** Håkon Hoel and Henrik Nyberg. *Gaussian Coarse Graining of a Master Equation Extension of Clarke's Model*, TRITA-NA 2012:5.
Submitted to Advances in Applied Probability.

The author contributed to all sections of the paper.

**Paper II** Håkon Hoel, Erik von Schwerin, Anders Szepessy, and Raúl Tempone. *Implementation and Analysis of an Adaptive Multilevel Monte Carlo Algorithm*, TRITA-NA 2012:6.
Shorter version published in Numerical Analysis of Multiscale Computations, Lecture Notes in CSE, Springer, cf. [20]. Full version submitted to Monte Carlo Methods and Applications.

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

**Paper III** Christian Bayer, Håkon Hoel, Erik von Schwerin, and Raúl Tempone. *On Non-Asymptotic Optimal Stopping Criteria in Monte Carlo Simulations*, TRITA-NA 2012:7.
In preparation.

The author contributed to all sections of the paper.

**Paper IV** Christian Bayer, Håkon Hoel, Ashraful Kadir, Petr Plecháč, Mattias Sandberg, Anders Szepessy, and Raúl Tempone. *How Accurate is Molecular Dynamics?* TRITA-NA 2012:8.
Earlier version available on Arxiv, cf. [2], full version to be submitted.

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

**Contents** **vii**

**I Introductory Chapters** **1**

**1 Introduction** **3**

**2 Probability background** **7**
2.1 Random Variables and Probability Measures . . . . . 7
2.2 Convergence of Random Variables and Limit Theorems for Sampling . . . . 11

**3 Monte Carlo Methods** **13**
3.1 Examples and Properties . . . . . 13
3.2 Variance Reduction . . . . . 15
3.3 Sequential Stopping Rules . . . . . 17

**4 Adaptive Weak Approximations for SDE** **19**
4.1 Stochastic Differential Equations (SDE) . . . . . 19
4.2 The Euler Method for Itô SDE . . . . . 22
4.3 Adaptive Weak Solution Approximation . . . . . 28

**5 Wireless Channel Modeling** **31**
5.1 The Multipath Fading Channel . . . . . 31
5.2 From MFC Models to Gaussian Processes . . . . . 36

**6 Classical and Quantum Mechanics** **39**
6.1 Classical Mechanics . . . . . 39
6.2 Quantum Mechanics . . . . . 43

**Bibliography** **51**

**II Included Papers** **55**

# 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

$$\int g(x)dP(x) \tag{1.0.1}$$

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

$$\left| \int g(x)dP(x) - \sum_{i=1}^{M} \frac{g(X_i)}{M} \right| \tag{1.0.2}$$

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

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

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 $\mathrm{E}[g(X_T)]$ with accuracy $\mathcal{O}(\mathrm{TOL})$ from $\mathcal{O}\big(\mathrm{TOL}^{-3}\big)$ by a single level method to $\mathcal{O}\big(\mathrm{TOL}^{-2}\log(\mathrm{TOL})^2\big)$ 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 $\mu = \mathrm{E}[X]$ using $M$ samples is the sample average

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

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

$$P\left(\left|\overline{X}M - \mu\right| > \mathrm{TOL}\right) \leq \delta, \tag{1.0.4}$$

for small, fixed accuracy-confidence constants $\mathrm{TOL} > 0$ and $\delta > 0$. Clearly, $P(\left|\overline{X}M - \mu\right| > \mathrm{TOL})$ 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 $\mathrm{E}[|X|^{\infty}] < C$, 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 $\mathrm{E}[X]$ 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

$$2\left(1 - \Phi\left(\frac{\sqrt{M}\mathrm{TOL}}{\overline{\sigma}_M}\right)\right) < \delta,$$

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.

## 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

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

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

$$\int_{\mathbb{R}^{3(N+n)}} g(\mathbf{X})\Phi(\mathbf{X},\mathbf{x})^*\Phi(\mathbf{X},\mathbf{x})\,d\mathbf{X}\,d\mathbf{x} - \lim_{T\to\infty}\frac{1}{T}\int_0^T g(\mathbf{X}(t))dt \tag{1.0.5}$$

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

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 Piere 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

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

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

[1] English translation (with additions) of *De Ratiociniis in Ludo Aleae* by Frans van Schooten and Christiaan Huygens (foreword, at least) in 1656.

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* $(\Omega, \mathfrak{F}, P)$ *with the* sample space $\Omega$*, the* event space $\mathfrak{F}$*, and the* probability measure $P : \mathfrak{F} \to [0, 1]$*. 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* $\mathfrak{F}$ *is a* $\sigma$-algebra*, i.e., it is a collection of subsets of* $\Omega$ *fulfilling the following:*

*(i)* $\mathfrak{F}$ *is non-empty.*

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

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

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

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

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

$$P\left(\cup_i A_i\right) = \sum_i P(A_i).$$

*(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* $(\Omega, \mathfrak{F}, P)$ *and an arbitrary measurable space* $(\Gamma, \mathfrak{G})$*, then* $X : \Omega \to \Gamma$ *is said to be a* random variable/vector *if the map from* $(\Omega, \mathfrak{F})$ *to* $(\Gamma, \mathfrak{G})$ *is measurable, i.e.,*

$$\{\omega : X(\omega) \leq G\} \in \mathfrak{F}, \quad \forall G \in \mathfrak{G}.$$

A frequently encountered type of r.v. is the real-valued r.v. mapping from $(\Omega, \mathfrak{F})$ to $(\mathbb{R}, \mathfrak{B})$ where $\mathfrak{B}$ 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

$$F(x) = \int_{-\infty}^{x} \mathrm{d}F(s) = \int_{-\infty}^{x} f(s)\,\mathrm{d}r.$$

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

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

and the *variance*

$$\sigma^2 = \mathrm{Var}(X) := \mathrm{E}\big[|X - \mathrm{E}[X]\,|^2\big]$$

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.*

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

*It has the probability space*

$$\Omega = \{Heads,\ Tails\}, \quad \mathfrak{F} = \{\emptyset, \{Heads\}, \{Tails\}, \Omega\},$$
$$\text{and } P(X = 1) = p = P(X = \{Tails\}),$$

*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*

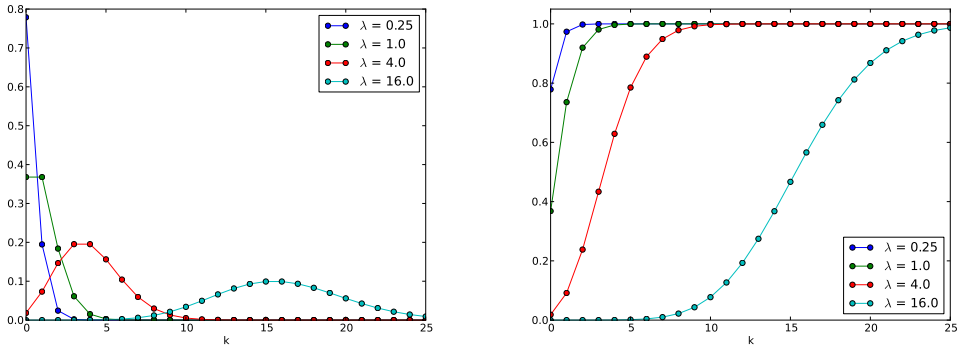


Figure 2.1: The Poisson PDF and CDF (both only defined at the dots for $k \in \mathbb{N}_0$—interpolating lines are included to illustrate the transitions).

*given by*

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

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

**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 $\Phi(x)$.*

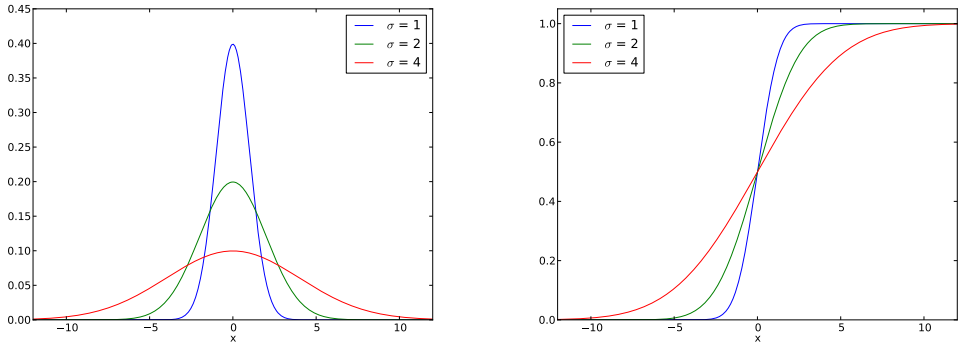


Figure 2.2: The PDFs and CDFs for normal distributions with $\mu = 0$.

*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*

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\det(K)^{1/2}} \exp\left(\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu)\right).$$

**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 $\mathrm{E}[X^n]$ for the Pareto r.v. only exists for $n < \alpha$ and, supposing $\alpha > 2$, its mean and variance are given by*

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

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

$$\lim_{x\to\infty} e^{\lambda x} P(X \geq x) = \infty, \quad \forall \lambda \in \mathbb{R}_+.$$

*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.*

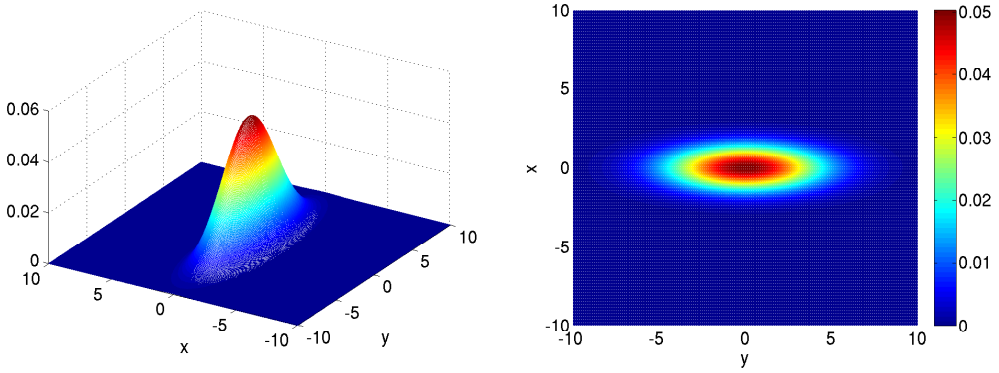


Figure 2.3: The PDF of the 2D multivariate normal distribution with $\mu = (0, 0)$ and the covariance matrix $K = \text{diag}(1, 10)$.

## 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

$$\textit{Almost sure conv.} \implies \textit{Conv. in probability} \implies \textit{Conv. in distribution.}$$

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 $(\Omega, \mathfrak{F})$ to $(\mathbb{R}, \mathfrak{B})$ are* independent *if for all $A_1, A_2 \in \mathfrak{B}$*

$$P(X_1 \in A_1, X_2 \in A_2) = P(X_1 \in A_1)P(X_2 \in A_2).$$

In particular, independence implies that $\mathrm{E}[X_1 X_2] = \mathrm{E}[X_1]\,\mathrm{E}[X_2]$. 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 $\mathrm{E}[|X_1|] < \infty$. 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 $\mathrm{Var}(X_1) < \infty$ using Chebycheff's inequality and independence: For any $\epsilon > 0$,

$$\begin{aligned} P(|S_n/n - \mu| > \epsilon) &\underbrace{\leq}_{\text{Chebycheff's ineq.}} \mathrm{E}\left[\frac{|S_n/n - \mu|^2}{\epsilon^2}\right] \\ &\leq \sum_{i=1}^{M} \frac{\mathrm{Var}(X_i)}{M^2\epsilon^2} \to 0 \text{ as } M \to \infty. \end{aligned}$$

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 $\mathrm{E}[|X_1|] < \infty$. 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) $\mu = \mathrm{E}[X_1]$ and $\sigma^2 = \mathrm{Var}(X_1) < \infty$. 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) $\mu = \mathrm{E}[X_1]$ and $\sigma^2 = \mathrm{Var}(X_1)$ and $\rho = \mathrm{E}\big[|X_1 - \mu|^3\big] < \infty$. 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

$$\overline{g}_M := \sum_{i=1}^{M} \frac{g(X_i)}{M}, \tag{3.0.1}$$

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 $\overline{g}_M$ is sufficiently small. MC methods are generally easy to implement, even for seemingly complicated problems, and their convergence rate is $\mathcal{O}\big(M^{-1/2}\big)$, 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.

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

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

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

with the accuracy-confidence constraint

$$P(\left|\overline{X}_M - \mu\right| > \text{TOL}) \leq \delta, \tag{3.1.2}$$

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

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

implies that to ensure that

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

we should use

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

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

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

So, if for example given the demands of accuracy $\text{TOL} = 0.02$ 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 $\sqrt{\mathrm{E}\left[\left|\overline{X}_M - \mu\right|^2\right]}$ follows from noting that for i.i.d. samples

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

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

$$\frac{\overline{X}_M - \mu}{\sqrt{M}} \sim \sigma\chi,$$

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

$$I = \int_A g(x)\,\mathrm{d}x \tag{3.1.5}$$

where $A$ is a compact subset of $\mathbb{R}^n$ and $g : A \to \mathbb{R}$. 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

$$\overline{I}_M = \sum_{i=1}^{M} \frac{g(X_i)}{M}. \tag{3.1.6}$$

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

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

equaling the expected convergence rate. In comparison, a cubature[1] approximation of (3.1.5) of order $k$ will have the convergence rate $\mathcal{O}\big(M^{-k/d}\big)$ 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.

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

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

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

$$\widetilde{I}_M := \sum_{i,j,k,l,m,n=1}^{\lceil M^{1/6} \rceil} g(x_i, x_j, x_k, x_l, x_m, x_n)h^6, \tag{3.1.7}$$

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

## 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

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

Generally, the cost of computing an MC estimate is

$$\mathcal{O}(M) = \mathcal{O}\left(\frac{\sigma^2 \left(\Phi^{-1}(1 - \delta/2)\right)^2}{\mathrm{TOL}^2}\right),$$

[1] Cubature is the name for numerical integration in dimensions higher than 1.

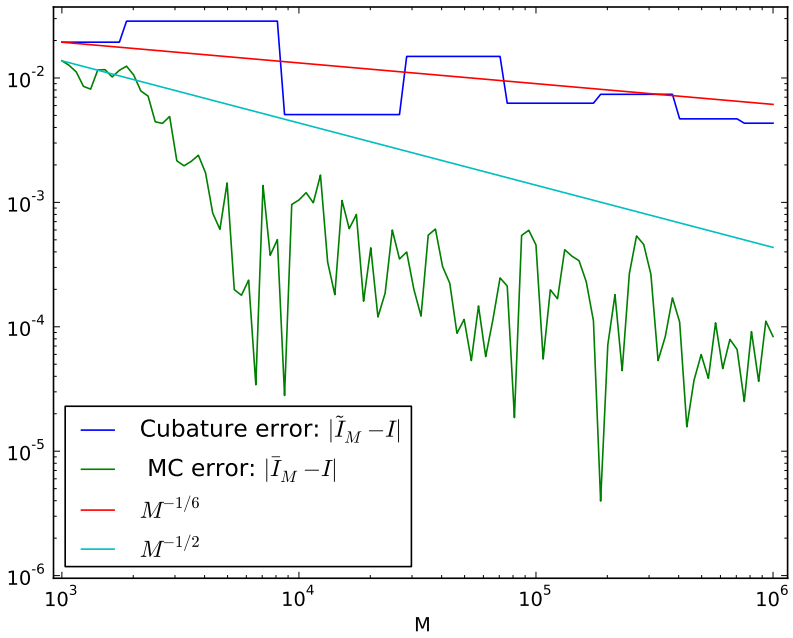


Figure 3.1: Comparison of convergence rates for approximating the volume of the unit ball in $\mathbb{R}^6$, $I = \pi^3/6$, when using either the MC algorithm (3.1.6) or the cubature (3.1.7). Due to the high dimensionality of the problem, the MC algorithm outperforms the first order cubature method.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

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 $\mathrm{E}[X]$ by a Monte Carlo method, say

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

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

$$\overline{Z}_M = \sum_{i=1}^{M} \frac{Z_i}{M}.$$

If the variance is reduced, $\mathrm{Var}(Z) = \mathrm{Var}(X - Y) < \mathrm{Var}(X)$, we reason from the introduction of this section that from a cost perspective, it is more efficient to approximate $\mathrm{E}[X]$ 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:** $\overline{X}_M$.

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

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

**while** $2\Big(1 - \Phi(\sqrt{M_k}\mathrm{TOL}/\overline{\sigma}_{M_k})\Big) > \delta$ **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 $\overline{\sigma}^2_{M_k}$ 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 $\overline{X}_M$.

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 $\overline{\sigma}^2_M$ used is positive and asymptotically consistent in the sense $\overline{\sigma}_M \to \sigma$ almost surely as $\mathrm{TOL} \to 0$, then for any confidence $\delta \in (0, 1)$, the stopping criterion

$$\frac{\overline{\sigma}^2_M}{M} \leq \frac{\mathrm{TOL}^2}{\left(\Phi^{-1}(1 - \delta/2)\right)^2} \tag{3.3.2}$$

implies the asymptotic fulfillment of the accuracy-confidence constraint

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

For practical purposes however, it is the reliability of sequential stopping rules for a fixed accuracy-confidence combination $\mathrm{TOL}, \delta \in (0, 1)$ 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 non-asymptotic regime is given by the Berry-Esseen Theorem 2.2.8 which presents the rate $\mathcal{O}\big(M^{-1/2}\big)$. But, the leading order constant of the Berry-Esseen bound contains the factor

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

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 $\mathrm{TOL}, \delta \in (0,1)$, 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:

$$\begin{aligned} dX &= a(t, X)dt + \text{“noise”}, \\ X_0 &= x_0, \end{aligned} \tag{4.1.1}$$

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

$$\begin{aligned} dX &= a(t, X)dt + b(t, X)dW_t, \\ X_0 &= x_0. \end{aligned} \tag{4.1.2}$$

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[1] $W : [0, T] \times \Omega \to \mathbb{R}$ which is characterized by the properties

[1] For more on Gaussian processes, see Chapter 5.

1. $W_0 = 0$,
2. The function $t \to W_t$ 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.

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

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

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

The Wiener process $W_t(\omega)$ is a function of two variables. For fixed $\omega \in \Omega$, we call $W_\cdot(\omega) : [0, T] \to \mathbb{R}$ 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) $W_t \in \mathcal{N}(0, t)$, that is, for any Borel set $B \subset \mathbb{R}$,

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

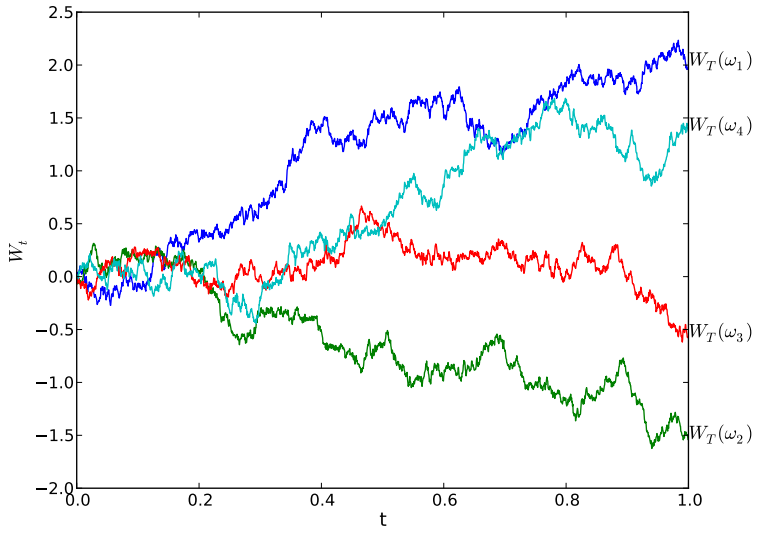


Figure 4.1: Four independent Wiener process realizations $W_t(\omega_i)$.

### 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

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

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

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

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 $\mathrm{E}[I_-(f)] = 0$, $\mathrm{E}[I_+(f)] = t$ 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

$$\int_0^t f(s,\omega)\, dW_s := \lim_{\Delta t \to 0} \sum_i f(t_i, \omega)\Delta W_i. \tag{4.1.5}$$

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

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

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"

$$dW_s^2 = dt, \quad dt \cdot dW_s = 0, \quad dt^2 = 0, \dots \tag{4.1.6}$$

yields following truncated Itô-Taylor expansion of $dg$

$$\begin{aligned} dg(t, X_t) &= \partial_t g(t, X_t)dt + \partial_X g(t, X_t)dX + \frac{1}{2}\partial_{XX} g(t, X_t)dX^2 \\ &= \left(\partial_t g(t, X_t) + a(t, X_t)\partial_X g(t, X_t) + \frac{(b(t, X_t))^2}{2}\partial_{XX} g(t, X_t)\right) dt \\ &\quad + b(t, X_t)\partial_X g(t, X_t)dW_t. \end{aligned} \tag{4.1.7}$$

Here we used (4.1.2) and the Itô "laws" (4.1.6) to derive that $dX^2 = b^2 dt + o(dt)$.

**Solving an SDE by Using Itô Calculus**

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

$$dX_t = rX_t dt + vX_t dW_t, \qquad r, v \in \mathbb{R}_+. \tag{4.1.8}$$

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

$$\frac{dX_t}{X_t} = rdt + vdW_t.$$

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).

## 4.2 The Euler Method for Itô SDE

We recall that a solution of an Itô SDE

$$\begin{aligned} dX &= a(t, X)dt + b(t, X)dW_t, \\ X_0 &= x_0. \end{aligned}$$

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

$$\int_0^t b(s, X_s)\, dW_s := \lim_{\Delta t \to 0} \sum_i b(t_i, X_{t_i})\Delta W_i.$$

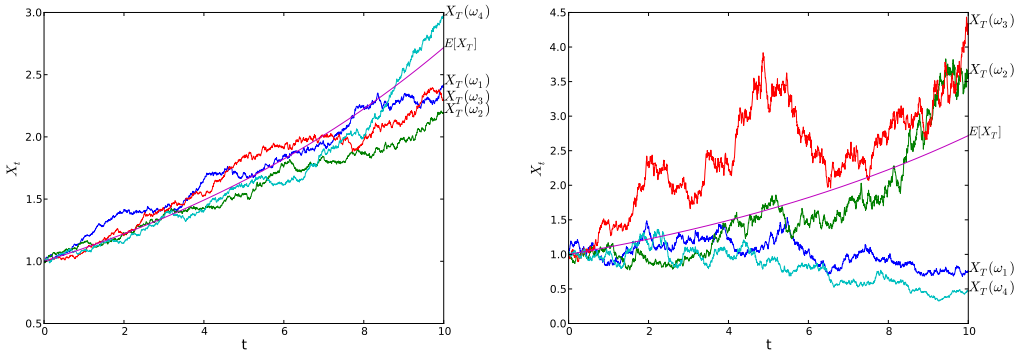


Figure 4.2: **Left plot:** Four independent sample path solutions of the SDE (4.1.8) when $X_0 = 1$, $r = 0.1$, and $v = 0.05$. **Right plot:** Four independent sample path solutions of the the SDE (4.1.8) when $X_0 = 1$, $r = 0.1$, and $v = 0.25$.

For numerical solutions, the (forward) Euler method generates realizations $\overline{X}_\cdot$ of the Itô SDE by the scheme

$$\begin{aligned} \overline{X}_{t_{n+1}} &= \overline{X}_{t_n} + a(t_n, \overline{X}_{t_n})\Delta t + b(t_n, \overline{X}_{t_n})\Delta W_n, \\ \overline{X}_0 &= x_0, \end{aligned} \tag{4.2.1}$$

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 $\mathrm{E}\big[|\overline{X}_T - X_T|\big]$ 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 $\mathrm{E}[g(X_t)]$ for some smooth function $g$. If so, the convergence criterion of interest is the *weak convergence*[2]

$$\sup_{g \in \mathbb{C}_C^\infty(\mathbb{R})} \big|\mathrm{E}[g(X_T)] - \mathrm{E}\big[g(\overline{X}_T)\big]\big|,$$

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 $\mathcal{O}\big(\Delta t^{1/2}\big)$, cf. [22, 8], and the weak convergence was shown by Talay and Tubaro to be $\mathcal{O}(\Delta t)$, 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*

[2]Weak convergence is equivalent to convergence in distribution, as defined in Chapter 2.

*expand to adaptive methods in higher dimensions. For more on higher order methods, we refer to [22].*

### 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

$$\begin{aligned} dX &= a(t,X)dt + b(t,X)dW_t, \quad 0 \le t \le T, \\ X_0 &= x_0, \quad x_0 \in \mathbb{R}. \end{aligned} \tag{4.2.2}$$

with an adaptive Euler method which fulfills the weak accuracy constraint

$$\left|\mathrm{E}\left[g(X_T) - g\left(\overline{X}_T\right)\right]\right| \le \mathrm{TOL}_\mathrm{T}, \tag{4.2.3}$$

where $X_t$ and $\overline{X}_t$ represents the explicit and numerical solution of (4.3.1), respectively, and $\mathrm{TOL}_\mathrm{T} > 0$ 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 $\overline{X}_t$ fulfill (4.2.3). As expected, the adaptive Euler method is given by

$$\overline{X}_{t_{n+1}} = \overline{X}_{t_n} + a(t_n, \overline{X}_{t_n})\Delta t_n + b(t_n, \overline{X}_{t_n})\Delta W_n, \tag{4.2.4}$$

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

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

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

$$\rho(t_n,\omega){\Delta t_n}^2(\omega) \le C\frac{\mathrm{TOL}_\mathrm{T}}{\mathrm{E}[N]}, \quad \forall n = 0,1,2,\ldots,N(\omega). \tag{4.2.6}$$

Given an initial time grid, time increments, and a Wiener process numerical sample path which we represent by $t_\cdot(\omega,0)$, $\Delta t_\cdot(\omega,0)$, and $W_{t_\cdot(\omega,0)}$, 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 $\rho(t_\cdot, \omega)$ for the numerical path represented by $t_\cdot(\omega, \ell)$, $\Delta t_\cdot(\omega, \ell)$, and $W_{t_\cdot(\omega,\ell)}$, cf. (4.2.9).
3. If condition (4.2.6) is fulfilled, then stop; otherwise
4. Refine the grid $t_\cdot(\omega, \ell)$ at all points where (4.2.6) is not fulfilled by halving the steps:

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

   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:

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

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

$$\overline{X}_{t_{n+1}} = c(t_n, \overline{X}_{t_n}), \text{ where } c(t_n, x) := x + a(t_n, x)\Delta t_n + b(t_n, \overline{X}_{t_n})\Delta W_n. \tag{4.2.8}$$

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

$$\begin{aligned} \rho(t_n, \cdot) = \frac{1}{2\Delta t_n} \Bigg( & \left(a(t_{n+1}, \overline{X}_{t_{n+1}}) - a(t_n, \overline{X}_{t_n})\right) \phi(t_{n+1}) \\ & + \left(b^2(t_{n+1}, \overline{X}_{t_{n+1}}) - b^2(t_n, \overline{X}_{t_n})\right) \partial_{\overline{X}_{t_{n+1}}} \phi(t_{n+1}) \Bigg), \end{aligned} \tag{4.2.9}$$

where $\phi$ is the solution of the discrete dual backward problem

$$\begin{aligned} \phi(t_n) &= \partial_x c(t_{n+1}, \overline{X}_{t_{n+1}})\phi(t_{n+1}), \quad t < T \\ \phi(T) &= g'(\overline{X}_T), \end{aligned} \tag{4.2.10}$$

and $\partial_{\overline{X}_{t_n}} \phi(t_n)$ is obtained by a longer scheme by linearising the forward problem (4.1.2), cf. [30, Theorem 2.2].

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

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

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

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

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

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

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

$$\begin{aligned} du(t,\overline{X}_t) &= \left(\partial_t u(t,\overline{X}_t) + \bar{a}(t;\overline{X})\partial_x u(t,\overline{X}_t) + \frac{\bar{b}^2(t;\overline{X})}{2}\partial_{xx}u(t,\overline{X}_t)\right) dt \\ &\quad + \bar{b}(t;\overline{X})\partial_x u(t,\overline{X}_t)\, dW_t \\ &= \left((\bar{a}(t;\overline{X}) - a(t,\overline{X}_t))\partial_x u(t,\overline{X}_t) + \frac{\bar{b}^2(t;\overline{X}) - b^2(t,\overline{X}_t)}{2}\partial_{xx}u(t,\overline{X}_t)\right) dt \\ &\quad + \bar{b}(t;\overline{X})\partial_x u(t,\overline{X}_t)\, dW_t. \end{aligned}$$

Recalling that $\overline{X}_0 = X_0$, we see that $u(0,\overline{X}_0) = \mathrm{E}[g(X_T)]$. Hence,

$$\begin{aligned} &\mathrm{E}[g(X_T)] - g(\overline{X}_T) = -\int_0^T du(t,\overline{X}_t) \\ &= \sum_n \int_{t_n}^{t_{n+1}} (a(t,\overline{X}_t) - a(t_n,\overline{X}_{t_n}))\partial_x u(t,\overline{X}_t) + \frac{b^2(t,\overline{X}_t) - b^2(t_n,\overline{X}_{t_n})}{2}\partial_{xx}u(t,\overline{X}_t)\, dt \\ &\quad - \sum_n \int_{t_n}^{t_{n+1}} b(t_n,\overline{X}_{t_n})\partial_x u(t,\overline{X}_t) dW_t. \end{aligned} \tag{4.2.12}$$

Taking the expected value of equation (4.2.12) yields

$$\begin{aligned} \mathrm{E}\big[g(X_T) - g(\overline{X}_T)\big] &= \sum_n \int_{t_n}^{t_{n+1}} \mathrm{E}\big[(a(t,\overline{X}_t) - a(t_n,\overline{X}_{t_n}))\partial_x u(t,\overline{X}_t)\big]\, dt \\ &\quad + \sum_n \int_{t_n}^{t_{n+1}} \mathrm{E}\left[\frac{b^2(t,\overline{X}_t) - b^2(t_n,\overline{X}_{t_n})}{2}\partial_{xx}u(t,\overline{X}_t)\right] dt, \end{aligned} \tag{4.2.13}$$

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 $\bar{u}(t,x) := \mathrm{E}\big[g(\overline{X}_T)|\overline{X}_t = x\big]$, the following approximations of the terms in (4.2.13) are valid:

$$\begin{aligned} &\int_{t_n}^{t_{n+1}} \mathrm{E}\big[a(t,\overline{X}_t) - (a(t_n,\overline{X}_{t_n}))\partial_x u(t,\overline{X}_t)\big]\, dt \\ &= \mathrm{E}\big[(a(t_{n+1},\overline{X}_{t_{n+1}}) - a(t_n,\overline{X}_{t_n}))\partial_x \bar{u}(t_{n+1},\overline{X}_{t_{n+1}})\big]\, \frac{\Delta t_n}{2} + \text{h.o.t.} \end{aligned}$$

and

$$\begin{aligned} &\int_{t_n}^{t_{n+1}} \mathrm{E}\left[\frac{b^2(t,\overline{X}_t) - b^2(t_n,\overline{X}_{t_n})}{2}\partial_{xx}u(t,\overline{X}_t)\right] dt \\ &= \mathrm{E}\left[\frac{b^2(t_{n+1},\overline{X}_{t_{n+1}}) - b^2(t_n,\overline{X}_{t_n})}{2}\partial_{xx}\bar{u}(t_{n+1},\overline{X}_{t_{n+1}})\right] \frac{\Delta t_n}{2} + \text{h.o.t.} \end{aligned}$$

Further, since

$$\partial_x \bar{u}(t,x) = \frac{\partial}{\partial x}\mathrm{E}\big[g(\overline{X}_T)|\overline{X}_t = x\big] = \mathrm{E}\Big[g'(\overline{X}_T)\partial_{\overline{X}_t}\overline{X}_T \,\Big|\, \overline{X}_t = x\Big],$$

we see that

$$\partial_x \bar{u}(t_{n+1}, \overline{X}_{t_{n+1}}) = \mathrm{E}\Big[g'(\overline{X}_T)\partial_{\overline{X}_{t_{n+1}}}\overline{X}_T \,\Big|\, \mathcal{F}_{t_{n+1}}\Big], \tag{4.2.14}$$

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

$$\begin{aligned}
0 &= \sum_{\ell=n}^{N-1} \big(\phi(t_\ell) - \partial_x c(\overline{X}_{t_\ell}, t)\phi(t_{\ell+1})\big)\,\partial_{\overline{X}_{t_n}}\overline{X}_{t_\ell} \\
&= \sum_{\ell=n}^{N-1} \phi(t_{\ell+1}) \underbrace{\Big(\partial_{\overline{X}_{t_n}}\overline{X}_{t_{\ell+1}} - \partial_x c(\overline{X}_{t_\ell}, t)\partial_{\overline{X}_{t_n}}\overline{X}_{t_\ell}\Big)}_{=0 \text{ cf. } (4.2.8)} - \phi(T)\frac{\partial \overline{X}_T}{\partial \overline{X}_{t_n}} + \phi(t_n)\frac{\partial \overline{X}_{t_n}}{\partial \overline{X}_{t_n}} \qquad (4.2.15)\\
&= -\phi(T)\frac{\partial \overline{X}_T}{\partial \overline{X}_{t_n}} + \phi(t_n).
\end{aligned}$$

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

$$\begin{aligned}
&\mathrm{E}\big[(a(t_{n+1}, \overline{X}_{t_{n+1}}) - a(t_n, \overline{X}_{t_n}))\partial_x \bar{u}(t_{n+1}, \overline{X}_{t_{n+1}})\big] \\
&= \mathrm{E}\big[(a(t_{n+1}, \overline{X}_{t_{n+1}}) - a(t_n, \overline{X}_{t_n}))\mathrm{E}\big[\phi(t_{n+1})\,|\,\mathcal{F}_{t_{n+1}}\big]\big] \qquad (4.2.16)\\
&= \mathrm{E}\big[(a(t_{n+1}, \overline{X}_{t_{n+1}}) - a(t_n, \overline{X}_{t_n}))\phi(t_{n+1})\big],
\end{aligned}$$

and by a similar argument it follows that

$$\begin{aligned}
&\mathrm{E}\left[\frac{b^2(t_{n+1}, \overline{X}_{t_{n+1}}) - b^2(t_n, \overline{X}_{t_n})}{2}\partial_{xx}\bar{u}(t_{n+1}, \overline{X}_{t_{n+1}})\right] \\
&= \mathrm{E}\left[\frac{b^2(t_{n+1}, \overline{X}_{t_{n+1}}) - b^2(t_n, \overline{X}_{t_n})}{2}\partial_{\overline{X}_{t_{n+1}}}\phi(t_{n+1})\right].
\end{aligned} \tag{4.2.17}$$

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*

$$\begin{aligned}
\rho(t_n,\omega) =& \frac{1}{2}\left(\partial_t a + a\partial_x a + \frac{b^2\partial_{xx}a}{2}\right)\phi(t_{n+1}) \\
&+ \left(\partial_t\frac{b^2}{2} + a\partial_x\frac{b^2}{2} + \frac{b^2}{2}\partial_{xx}\frac{b^2}{2} + b^2\partial_x a\right)\partial_{\overline{X}_{t_{n+1}}}\phi(t_{n+1}) \\
&+ \frac{b^2}{2}\partial_{xx}b^2\frac{\partial^2\phi(t_{n+1})}{\partial\overline{X}^2_{t_{n+1}}},
\end{aligned}$$

*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

$$\begin{aligned} dX &= a(t,X)dt + b(t,X)dW_t, \quad 0 \le t \le T, \\ X_0 &= x, \end{aligned} \tag{4.3.1}$$

we consider the problem of minimizing the computational cost of approximating $\mathrm{E}[g(X(T)]$ within a given tolerance $\mathrm{TOL} > 0$. 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.

### Single Level Weak Approximation

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

$$\mathcal{A}\big(g(\overline{X}_T); M\big) = \sum_{i=1}^{M} \frac{g(\overline{X}_T(\omega_i))}{M}. \tag{4.3.2}$$

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

$$\begin{aligned} \big|\mathrm{E}[g(X_T)] - \mathcal{A}\big(g(\overline{X}_T); M\big)\big| \le & \big|\mathrm{E}[g(X_T)] - \mathrm{E}\big[g(\overline{X}_T)\big]\big| \\ & + |E[g(\overline{X}(T))] - \mathcal{A}\big(g(\overline{X}_T); M\big)| =: \mathcal{E}_T + \mathcal{E}_S, \end{aligned}$$

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 $\mathcal{E}_S = \mathcal{O}\big(M^{-1/2}\big)$. So to ensure that $|E[g(X(T))] - \mathcal{A}\big(g(\overline{X}_T); M\big)| = \mathcal{O}(\mathrm{TOL})$, one should choose $\Delta t = \mathcal{O}(\mathrm{TOL})$ and $M = \mathcal{O}\big(\mathrm{TOL}^{-2}\big)$. The computational cost thus becomes $\mathcal{O}\big(\mathrm{TOL}^{-3}\big)$ 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 $\overline{X}_t^{(\ell)}$ denote an Euler method realization on a grid with uniform step size $\Delta t^{(\ell)}$. Then the MLMC method approximates $E[g(X(T))]$ by the telescoping sum

$$\mathcal{A}_{\mathcal{ML}}\big(g(\overline{X}_T); M_0\big) = \sum_{i=1}^{M_0} \frac{g\Big(\overline{X}_T^{(0)}(\omega_{i,0})\Big)}{M_0} + \sum_{\ell=1}^{L}\sum_{i=1}^{M_\ell} \frac{g\Big(\overline{X}_T^{(\ell)}(\omega_{i,\ell})\Big) - g\Big(\overline{X}_T^{(\ell-1)}(\omega_{i,\ell})\Big)}{M_\ell}, \quad (4.3.3)$$

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 $\overline{X}_T^{(\ell)}(\omega_{i,\ell})$ and $\overline{X}_T^{(\ell-1)}(\omega_{i,\ell})$ 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

$$\begin{aligned} W_{t_{2n-1}^{(\ell)}}(\omega) &= W_{t_n^{(\ell-1)}}(\omega) \\ W_{t_{2n}^{(\ell)}}(\omega) &= \frac{W_{t_n^{(\ell-1)}}(\omega) + W_{t_{n+1}^{(\ell-1)}}(\omega)}{2} + \frac{\sqrt{\Delta t^{(\ell-1)}}}{2}\xi, \quad \text{where} \quad \xi \sim \mathcal{N}(0,1), \end{aligned}$$

cf. Figure 4.3.

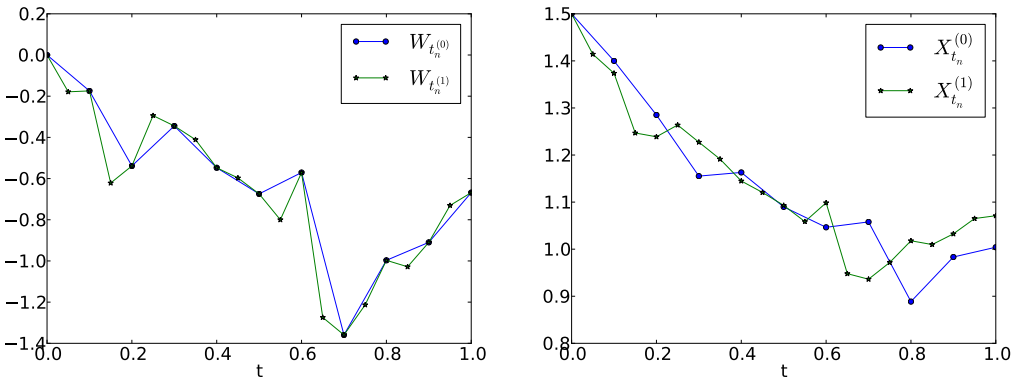


Figure 4.3: **Left plot:** A sample path $W_t(\omega)$ plotted on the coarse grid $W_{t_n^{(0)}}(\omega)$ (blue line ) with $\Delta t^{(0)} = 1/10$, and its finer level pair generated by Brownian bridges, $W_{t_n^{(1)}}(\omega)$ (green line) with $\Delta t^{(1)} = \Delta t^{(0)}/2$. **Right plot:** Euler 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

$$\mathrm{E}\big[\mathcal{A}_{\mathcal{ML}}\big(g(\overline{X}_T); M_0\big)\big] = \mathrm{E}\Big[g\Big(\overline{X}_T^{(L)}\Big)\Big],$$

and since sample paths are i.i.d.

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

The strong convergence $\mathrm{E}\Big[\Big(g(\overline{X}_T^{(\ell)}) - g(X_T)\Big)^2\Big] = \mathcal{O}\big(\Delta t^{(\ell)}\big)$, cf. Section 4.2, then further implies that

$$\begin{aligned}\mathrm{Var}\big(\mathcal{A}_{\mathcal{ML}}(g(\overline{X}_T); M_0)\big) &= \frac{\mathrm{Var}\Big(g\Big(\overline{X}_T^{(0)}\Big)\Big)}{M_0} + \sum_{\ell=1}^{L} \frac{\mathrm{Var}\Big(g\Big(\overline{X}_T^{(\ell)}\Big) - g\Big(\overline{X}_T^{(\ell-1)}\Big)\Big)}{M_\ell} \\ &= \mathcal{O}\left(M_0^{-1} + \sum_{\ell=1}^{L} 2^{-\ell} M_\ell^{-1}\right) = \mathcal{O}\big(LM_0^{-1}\big)\,.\end{aligned} \tag{4.3.4}$$

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

$$\big|\mathrm{E}[g(X_T)] - \mathcal{A}_{\mathcal{ML}}(g(\overline{X}_T); M_0)\big| \leq \mathcal{O}(\mathrm{TOL})$$

at the computational cost

$$\mathcal{O}\left(\sum_{\ell=0}^{L} \frac{M_\ell}{\Delta t^{(\ell)}}\right) = \mathcal{O}\Big(\big(\log(\mathrm{TOL}^{-1})\big)^2\,\mathrm{TOL}^{-2}\Big)\,.$$

This might be shown by splitting MLMC error into two contributions

$$\begin{aligned}\big|\mathrm{E}[g(X_T)] - \mathcal{A}_{\mathcal{ML}}(g(\overline{X}_T); M_0)\big| \leq &\Big|\mathrm{E}\Big[g(X_T) - g\Big(\overline{X}_T^{(L)}\Big)\Big]\Big| \\ &+ \Big|\mathrm{E}\Big[g\Big(\overline{X}_T^{(L)}\Big)\Big] - \mathcal{A}_{\mathcal{ML}}(g(\overline{X}_T); M_0)\Big| =: \mathcal{E}_T + \mathcal{E}_S.\end{aligned}$$

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

With Giles' MLMC method, the computational cost is thus improved from $\mathcal{O}\big(\mathrm{TOL}^{-3}\big)$ for the single level MC method to $\mathcal{O}\big((\log(\mathrm{TOL}^{-1}))^2\mathrm{TOL}^{-2}\big)$. 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 $\mathcal{A}_{\mathcal{ML}}\big(g(\overline{X}_T); M_0\big)$ has lower variance than the single level average operator $\mathcal{A}\big(g(\overline{X}_T); M\big)$, 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* $u(t,x) = \mathrm{E}[X_T|X_t = x]$, *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

$$Z_{t,M} = \sum_{j=1}^{M} a(\alpha_j) e^{-i2\pi f_c \tau(\alpha_j,t)} X_{t-\tau(\alpha_j,t)}, \tag{5.1.1}$$

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 $\mathcal{F}(X_\cdot)(f) \in C([-W, W])$. 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

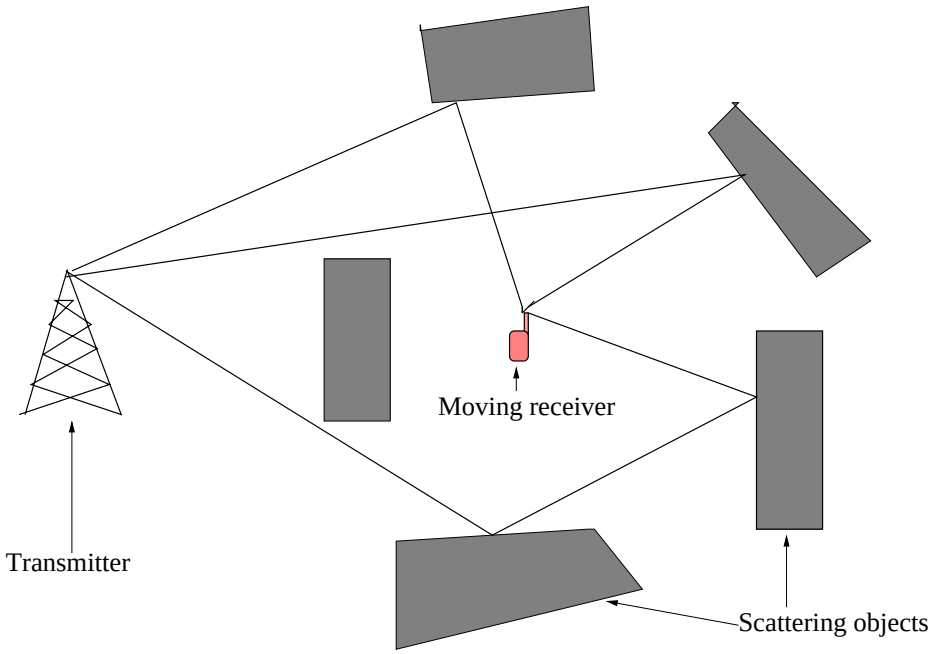


Figure 5.1: Illustration of typical scattering scenario considered for the MFC model. Lines between the transmitter, scatterers and the receiver represent different radio wave paths.

$X_t$ is modulated to the carrier frequency $f_c$ by the procedure

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

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

$$\widetilde{Z}_t = \int_{\mathbb{R}} \widetilde{H}(s,t)\widetilde{X}_{t-s}\, ds \tag{5.1.2}$$

where the passband channel response for MFC channels is given by

$$\widetilde{H}(s,t) := \sum_{j=1}^{M} a(\alpha_j)\delta(s-\tau(\alpha_j,t)), \tag{5.1.3}$$

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*

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

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

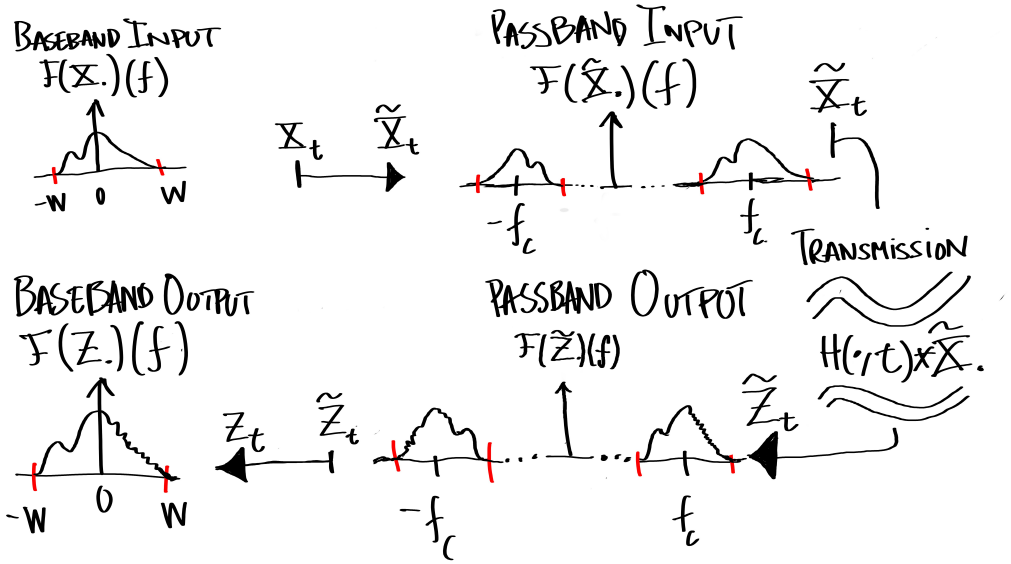


Figure 5.2: The modulation and demodulation taking place when transmitting data in a frequency modulated wireless channel. The baseband input is not drawn equal to the baseband output in the figure, i.e., $\mathcal{F}(Z_\cdot) \neq \mathcal{F}(X_\cdot)$; this is to visualize that some signal information is lost during transmission.

baseband channel response

$$H(s,t) := \sum_{j=1}^{M} a(\alpha_j) e^{-i2\pi f_c \tau(\alpha_j,t)} \delta(s - \tau(\alpha_j,t)), \tag{5.1.4}$$

it is possible to represent the output baseband as a function of the input baseband

$$Z_t = \int_{\mathbb{R}} H(s,t) X_{t-s}\, ds, \tag{5.1.5}$$

cf. [32]. This representation equals (5.1.1), avoids dealing with the passband, and it shows that given the baseband input signal, the 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 $a = 1/\sqrt{M}$;

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

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* $A_M(t) := \mathrm{E}\big[Z_{t,M} Z^*_{0,M}\big]$ 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*

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

**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 $p(\alpha) = (2\pi)^{-1}$, 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

$$S(f) = \begin{cases} \frac{c}{\pi\sqrt{(vf_c)^2 - (cf)^2}} & |f| < v f_c/c \\ 0 & |f| \geq v f_c/c, \end{cases} \tag{5.1.7}$$

cf. Figure 5.3.

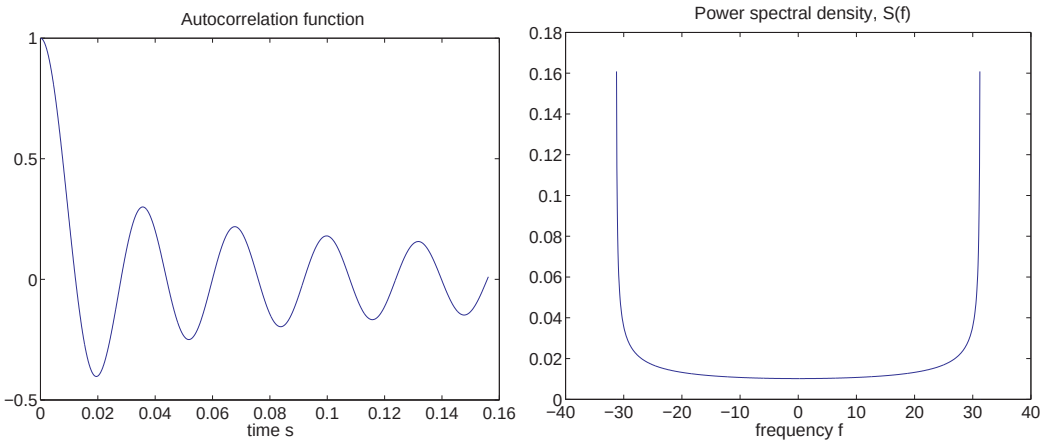


Figure 5.3: **Left plot:** The autocorrelation function for Clarke's model with azimuth density $p(\alpha) = (2\pi)^{-1}$, $v = 5m/s$ and $f_c = 1.8775GHz$. **Right plot:** The power spectral density of Clarke's model, often called Jakes' spectrum with the same model conditions as for the left plot.

### 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 atcive 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

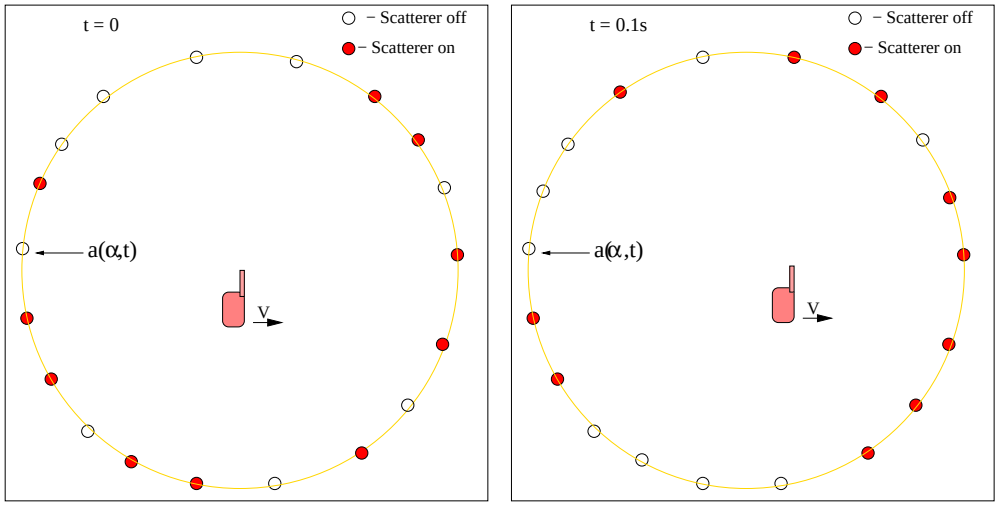


Figure 5.4: Moving receiver in the center of a circular scattering environment with scatterers flipping on and off as time goes.

$a^+ : \Omega \to \mathbb{R}_+$ 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 $C$:

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

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

$$Z_{t,M} = \frac{1}{\sqrt{M}} \sum_{m=1}^{M} a(\alpha_m, t) e^{-i2\pi f_c v \cos(\alpha_m) t/c + \theta_m(t)}. \tag{5.1.8}$$

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.

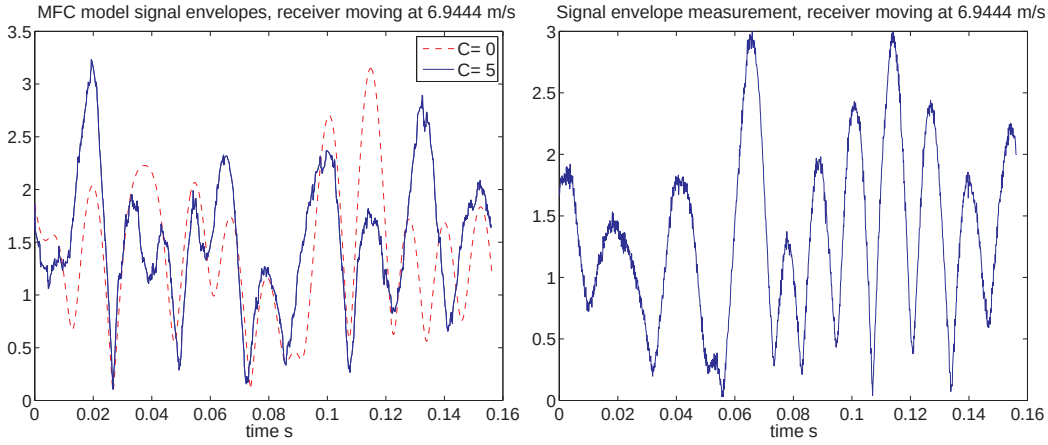


Figure 5.5: **Left plot:** Two computer generated signal realizations of the MFC model generated with same random seed initialization. Both realizations are generated with the modeling parameters $f_c = 1.8775GHz$, $v = 6.944m/s$, $a^+(\alpha) = 2/\sqrt{M}$, but the flip rate separates the realizations 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* $\mathbf{Z} = (Z_{t_1}, Z_{t_2}, \ldots, Z_{t_n})$ *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

$$Z_{t_i} = \sum_{j=1}^{N} \sqrt{K}_{i,j} \chi_j, \tag{5.2.1}$$

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*

$$K_{i,j} = E[W_{t_i} W_{t_j}] = E[W^2_{\min(t_i,t_j)}] = \min(t_i, t_j), \tag{5.2.2}$$

*cf. Section 4.1. For the Wiener process, the structure of the Cholesky factorized* $\sqrt{K}$ *is particularly simple:*

$$\sqrt{K} = \begin{pmatrix} \sqrt{t_1} & 0 & \ldots & 0 \\ \sqrt{t_1} & \sqrt{t_2 - t_1} & \ddots & 0 \\ \vdots & \vdots & \ddots & 0 \\ \sqrt{t_1} & \sqrt{t_2 - t_1} & \ldots & \sqrt{t_N - t_{N-1}} \end{pmatrix}. \tag{5.2.3}$$

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

$$W_{t_{j+1}} = W_{t_j} + \sqrt{t_{j+1} - t_j}\chi_j, \qquad \chi_j \sim \mathcal{N}(0,1).$$

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; $\mathbf{F} = -\nabla V$. This yields the equations of motion

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

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

$$\dot{E} = M\dot{\mathbf{v}} \cdot \mathbf{v} + \nabla V \cdot \dot{\mathbf{q}} \overbrace{=}^{(6.1.1)} 0. \tag{6.1.2}$$

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

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

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

For a system of $N$ particles with the $i$th particle position denoted $\mathbf{q}_i \in \mathbb{R}^d$ 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 $V_{ij}(\mathbf{q}_i - \mathbf{q}_j)$. Introducing the Hamiltonian

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

where $\mathbf{q} = (\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_N) \in \mathbb{R}^{dN}$ and, similarly, $\mathbf{p} \in \mathbb{R}^{dN}$, the equations of motion for $N$ particles take the familiar form

$$\dot{\mathbf{q}}_i = \nabla_{\mathbf{p}_i} H \quad \text{and} \quad \dot{\mathbf{p}}_i = -\nabla_{\mathbf{q}_i} H, \qquad \text{for} \quad i = 1, 2, \ldots, N. \tag{6.1.4}$$

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 $\mathbf{F} = -Mkq$, with $k \in \mathbb{R}_+$ being a spring constant and $q \in \mathbb{R}$ the displacement from equilibrium. The Hamiltonian for this dynamics becomes*

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

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

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

*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*

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

### Evolution of States

For a given $N$-particle system, the set of possible $(\mathbf{q}, \mathbf{p})$ restricted to some subset of $\mathbb{R}^{dN} \times \mathbb{R}^{dN}$ is referred to as the phase space, and the vector $(\mathbf{q}(t), \mathbf{p}(t))$ 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 $\rho(\mathbf{q}, \mathbf{p})$ with the restriction that $\rho$ is non-negative. A pure state $(\tilde{\mathbf{q}}, \tilde{\mathbf{p}})$ may then be represented on density form by dirac functions $\rho(\mathbf{q}, \mathbf{p}) = \delta(\mathbf{q} - \tilde{\mathbf{q}}, \mathbf{p} - \tilde{\mathbf{p}})$. For a given state $\rho$, the expected value of an observable, like the total energy $H$, is given by

$$E_\rho[H] = \frac{\int H(\mathbf{q}, \mathbf{p})\rho(\mathbf{q}, \mathbf{p})\, d\mathbf{q}\, d\mathbf{p}}{\int \rho(\mathbf{q}, \mathbf{p})\, d\mathbf{q}\, d\mathbf{q}}.$$

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

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

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

$$\frac{d}{dt}\int_{R(t)} \rho(\mathbf{q}, \mathbf{p}; t)\, d\mathbf{q}\, d\mathbf{q} = 0,$$

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

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

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

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

Furthermore, the Hamiltonian dynamics (6.1.4) implies that

$$\nabla_{\mathbf{q},\mathbf{p}} \cdot (\dot{\mathbf{q}}, \dot{\mathbf{p}}) = \sum_{i=1}^{N} (\nabla_{\mathbf{q}_i} \cdot \nabla_{\mathbf{p}_i} - \nabla_{\mathbf{p}_i} \cdot \nabla_{\mathbf{q}_i})\, H = 0, \tag{6.1.7}$$

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

$$\partial_t \rho + \sum_{i=1}^{N} \nabla_{\mathbf{p}_i} H \cdot \nabla_{\mathbf{q}_i} \rho - \nabla_{\mathbf{q}_i} H \cdot \nabla_{\mathbf{p}_i} \rho = 0 \tag{6.1.8}$$

with initial condition $\rho(\mathbf{q}, \mathbf{p}; 0) = \rho(\mathbf{q}, \mathbf{p})$. 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 $(\dot{\mathbf{q}}, \dot{\mathbf{p}})$ with the evolving phase space volume

$$V(t) = \int_{R(t)} 1\, d\mathbf{q}\, d\mathbf{p}.$$

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

$$\dot{V}(t) = \int_{R(t)} \nabla_{\mathbf{q},\mathbf{p}} \cdot (\dot{\mathbf{q}}, \dot{\mathbf{p}})\, d\mathbf{q}\, d\mathbf{p} = 0.$$

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

### Symplectic Numerical Methods

The word symplectic derives from the Ancient Greek $\sigma\upsilon\mu\pi\lambda\varepsilon\kappa\tau\iota\kappa\acute{o}\sigma$ 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

$$\sum_{i=1}^{n} (\xi_i^p)^* \eta_i^q - (\xi_i^q)^* \eta_i^p, \tag{6.1.9}$$

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*

$$A^T JA = J$$

**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 $g'(\mathbf{q}, \mathbf{p})$ is everywhere symplectic, i.e., if*

$$g'(\mathbf{q}, \mathbf{p})^T J g'(\mathbf{q}, \mathbf{p}) = J.$$

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

$$\phi_t'(\mathbf{q}, \mathbf{p})^T J \phi_t'(\mathbf{q}, \mathbf{p}) = J.$$

By observing that $\det(J) = (-1)^{2n} = 1$, $\det(\phi_0'(\mathbf{q}, \mathbf{p})) = 1$ and assuming the Hamiltonian is sufficiently smooth to make $\phi_t$ a continuous function with respect to $t$, we conclude that $\det(\phi_t'(\mathbf{q}, \mathbf{p})) = 1$ 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,*

$$\left(\frac{\partial(p_{n+1}, q_{n+1})}{\partial(p_n, q_n)}\right)^T J \left(\frac{\partial(p_{n+1}, q_{n+1})}{\partial(p_n, q_n)}\right) = J.$$

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

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

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*

$$(q_{n+1}, p_{n+1}) = (q_n + \Delta t p_n,\, p_n - \Delta t q_n)\,, \qquad \textit{(forward Euler)},$$
$$(q_{n+1}, p_{n+1}) = (q_n + \Delta t(p_n - \Delta t q_n),\, p_n - \Delta t q_n)\,, \qquad \textit{(sympltectic Euler)}.$$

*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.*

## 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
$$i\partial_t \Phi = \left(-\frac{1}{2M}\Delta + V(x, t)\right)\Phi.$$
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 $E = |\mathbf{p}|^2/2m + V(\mathbf{x}, t)$ with $\mathbf{p}, \mathbf{x} \in \mathbb{R}^3$. 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

$$\Phi(\mathbf{x}, t) = \phi(\mathbf{x}, t)e^{i(\mathbf{k}\cdot\mathbf{x} - 2\pi f t)} \tag{6.2.1}$$

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

$$\mathbf{p} = h\mathbf{k}, \quad |\mathbf{k}| = 2\pi f/c, \quad \text{and} \qquad E = |\mathbf{p}|c = h|\mathbf{k}|c = \hbar f. \tag{6.2.2}$$

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

$$\Phi(\mathbf{x}, t) = \phi(\mathbf{x}, t)e^{i(\mathbf{p}\cdot\mathbf{x} - Et)/\hbar}. \tag{6.2.3}$$

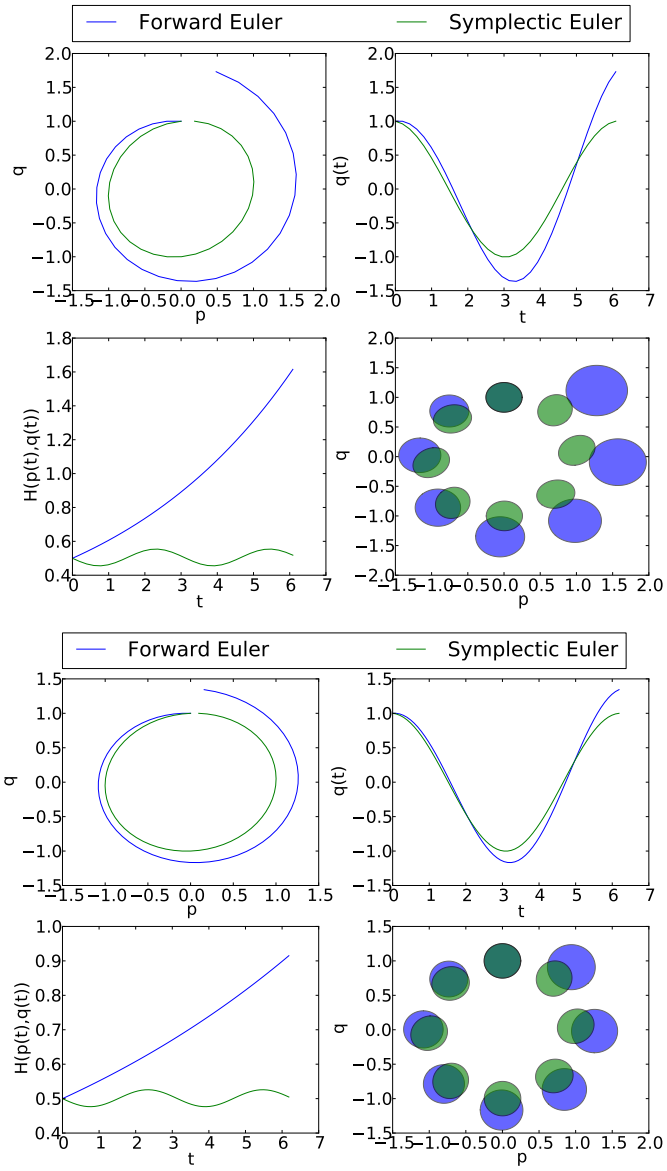


Figure 6.1: Numerical solutions of the Harmonic oscillator problem of Example 6.1.6. The top four plots are numerical solutions with $\Delta t = 2\pi/32 \approx 0.2$ integrated on the interval $t \in [0, 2\pi)$ and consist of: phase space solution $(p(t), q(t))$ (upper left), $(t, q(t))$(upper right),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

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

$$(i\hbar\partial_t + \hbar^2 \frac{\Delta_x}{2m})\Phi = (E - \frac{|\mathbf{p}|^2}{2m})\Phi. \tag{6.2.4}$$

By coordinate transformations $t \to t/\hbar, x \to x/\hbar$ and replacing $E$ with the classical mechanics total energy $|\mathbf{p}|^2/2m+V(\mathbf{x},t)$, equation (6.2.4) becomes the time-dependent Schrödinger equation

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

### 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

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

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

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

where $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N) \in \mathbb{R}^{3N}$ and $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) \in \mathbb{R}^{3n}$ denote the nuclear and electron positions, respectively, $M_i$ denotes the mass of the $i$-th nucleus, and $m \approx 9.11 \times 10^{-31}$kg the electron mass. Supposing all nuclei have the same mass, i.e., $M_i = M_1$ $\forall i$, and introducing the coordinate transformations $(\mathbf{X},\mathbf{x}) \to \sqrt{m}(\mathbf{X},\mathbf{x})$ and the short hand notation $\Delta_{\mathbf{X}} := \sum_{i=1}^{N}\Delta_{\mathbf{X}_i}$, we obtain the time-independent Schrödinger equation on the form studied in paper IV:

$$\left(-\frac{1}{2M}\Delta_{\mathbf{X}} \underbrace{-\sum_{i=1}^{n}\frac{1}{2}\Delta_{\mathbf{x}_i} + V(\mathbf{X},\mathbf{x})}_{=:\mathcal{V}(\mathbf{X},\mathbf{x})}\right)\Phi(\mathbf{X},\mathbf{x}) = E\Phi(\mathbf{X},\mathbf{x}), \tag{6.2.5}$$

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

[1] Isotope refers to the number of protons and neutrons in the nucleus of a chemical element.

### 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

$$\iint |\Phi(\mathbf{X}, \mathbf{x})|^2 \, d\mathbf{X} \, d\mathbf{x} = 1,$$

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

$$E_\Phi[\mathcal{A}] = \iint \Phi(\mathbf{X}, \mathbf{x})^* \mathcal{A}(\mathbf{X}) \Phi(\mathbf{X}, \mathbf{x}) \, d\mathbf{X} \, d\mathbf{x}.$$

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

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,

$$\left( -\sum_{i=1}^{N} \frac{1}{2M} \Delta_{\mathbf{X}_i} - \mathcal{V} \right) (a_1 \Phi_1 + a_2 \Phi_2) = E(a_1 \Phi_1 + a_2 \Phi_2), \qquad \forall a_1, a_2 \in \mathbb{C}$$

### 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

$$\Phi(\mathbf{X}, \mathbf{x}) = \phi(\mathbf{X}, \mathbf{x}) e^{i\sqrt{M}\theta(\mathbf{X})}, \tag{6.2.6}$$

where the amplitude $\phi : \mathbb{R}^{3N} \times \mathbb{R}^{3n} \to \mathbb{C}$ and the phase $\theta : \mathbb{R}^{3N} \to \mathbb{R}$ 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

$$\begin{aligned} 0 &= \left( -\frac{1}{2} \Delta_{\mathbf{X}} + \mathcal{V}(\mathbf{X}, \mathbf{x}) - E \right) \phi(\mathbf{X}, \mathbf{x}) e^{i\sqrt{M}\theta(\mathbf{X})} \\ &= \Bigg( \underbrace{(\frac{1}{2}|\nabla\theta|^2 + V_0 - E)\phi}_{=:I} \\ &\qquad \underbrace{-\frac{1}{2M}\Delta_{\mathbf{X}}\phi + (\mathcal{V} - V_0)\phi - \frac{i}{M^{1/2}}(\nabla_{\mathbf{X}}\phi \cdot \nabla_{\mathbf{X}}\theta + \frac{1}{2}\phi \, \Delta_{\mathbf{X}}\theta)}_{=:II} \Bigg) e^{i\sqrt{M}\theta(\mathbf{X})}. \end{aligned} \tag{6.2.7}$$

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 $M \to \infty$ (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

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

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

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

and $G : \mathbb{R}^{3N} \to \mathbb{R}$ 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

$$\rho_{\mathrm{S}}(\mathbf{X}) = \frac{G^{-2}(\mathbf{X})\langle\psi(\mathbf{X},\cdot),\psi(\mathbf{X},\cdot)\rangle}{\|G^{-1}\psi\|_{L^2(\mathbb{R}^{3N}\times\mathbb{R}^{3n})}}.$$

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

$$\mathcal{V}(\mathbf{X},\cdot)\psi_0(\mathbf{X},\cdot) = \lambda_0(\mathbf{X})\psi_0(\mathbf{X},\cdot). \tag{6.2.8}$$

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

[2]Often referred to as a potential energy surface.

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

$$\left(\frac{1}{2M}\Delta_{\mathbf{X}} + \lambda_0(\mathbf{X})\right)\Phi(\mathbf{X}) = E\Phi(\mathbf{X}). \tag{6.2.9}$$

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

$$\Phi_{\mathrm{BO}}(\mathbf{X},\mathbf{x}) = \phi_{\mathrm{BO}}(\mathbf{X},\mathbf{x})e^{i\sqrt{M}\theta_{\mathrm{BO}}(\mathbf{X})},$$

where, as before, we assume that the amplitude $\phi_{\mathrm{BO}} : \mathbb{R}^{3N}\times\mathbb{R}^{3n}\to\mathbb{C}$ and phase $\theta_{\mathrm{BO}} : \mathbb{R}^{3N}\to\mathbb{R}$ 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 $\psi_0(\mathbf{X},\cdot)$ of (6.2.8) with eigenfunction $\lambda_0(\mathbf{X})$ so that the amplitude takes the form $\phi_{\mathrm{BO}}(\mathbf{X},\mathbf{x}) = \sqrt{\rho_{\mathrm{BO}}}(\mathbf{X})\psi_0(\mathbf{X},\mathbf{x})$ with unknown density function $\rho_{\mathrm{BO}} : \mathbb{R}^{3N}\to\mathbb{R}$. Inserting $\Phi_{\mathrm{BO}}$ into equation (6.2.7), yields

$$\begin{aligned}
0 &= \left(\frac{1}{2M}\Delta_{\mathbf{X}} + \lambda_0(\mathbf{X}) - E\right)\Phi_{\mathrm{BO}} \\
&= \Bigg(\underbrace{(\frac{1}{2}|\nabla\theta_{\mathrm{BO}}|^2 + \lambda_0 - E)}_{=:I}\phi_{\mathrm{BO}} \\
&\qquad -\frac{1}{2M}\Delta_{\mathbf{X}}\phi_{\mathrm{BO}} - \frac{i}{M^{1/2}}(\nabla_{\mathbf{X}}\phi_{\mathrm{BO}}\cdot\nabla\theta_{\mathrm{BO}} + \frac{1}{2}\phi_{\mathrm{BO}}\,\Delta\theta_{\mathrm{BO}})\Bigg)e^{i\sqrt{M}\theta_{\mathrm{BO}}(\mathbf{X})}.
\end{aligned} \tag{6.2.10}$$

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

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

whose characteristics are given by the Hamiltonian system

$$\begin{aligned}
&\dot{\mathbf{X}}_i = \partial_{\mathbf{P}_i}H_{\mathrm{BO}}(\mathbf{X},\mathbf{P}), \qquad\qquad \dot{\mathbf{P}} = -\partial_{\mathbf{X}_i}H_{\mathrm{BO}}(\mathbf{X},\mathbf{P}), \\
&\text{where}\quad H_{\mathrm{BO}}(\mathbf{X},\mathbf{P}) := \sum_{i=1}^{N}\frac{|\mathbf{P}_i|^2}{2M} + \lambda_0(\mathbf{X}), \quad \text{and} \quad \mathbf{P}(t) = \nabla_{\mathbf{X}}\theta_{\mathrm{BO}}(\mathbf{X}(t)).
\end{aligned} \tag{6.2.11}$$

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

$$\nabla\cdot(\rho_{\mathrm{BO}}(\mathbf{X})\nabla\theta_{\mathrm{BO}}(\mathbf{X})) = 0.$$

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

$$\frac{d}{dt}\rho_{\mathrm{BO}}(\mathbf{X}(t)) = \nabla\rho_{\mathrm{BO}}(\mathbf{X}(t))\cdot\nabla\theta_{\mathrm{BO}}(\mathbf{X}(t)) = -\rho_{\mathrm{BO}}(\mathbf{X}(t))\Delta\theta_{\mathrm{BO}}(\mathbf{X}(t)).$$

Introducing the integrating factor

$$\frac{d}{dt}\log G_{\mathrm{BO}}(\mathbf{X}(t)) = \frac{1}{2}\Delta\theta_{\mathrm{BO}}(\mathbf{X}(t)),$$

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

$$\rho_{\mathrm{BO}}(\mathbf{X}) = \frac{C}{G^2_{\mathrm{BO}}(\mathbf{X})}.$$

The main study of Paper IV is the comparison of the densities

$$\rho_S(\mathbf{X}) = \frac{G^{-2}(\mathbf{X})\langle\psi(\mathbf{X},\cdot),\psi(\mathbf{X},\cdot)\rangle}{\|G^{-1}\psi\|_{L^2(\mathbb{R}^{3N}\times\mathbb{R}^{3n})}} \text{ and } \rho_{\mathrm{BO}}(\mathbf{X}) = \frac{C}{G^2_{\mathrm{BO}}(\mathbf{X})},$$

and we end this chapter by noting that Theorem 7.1 of Paper IV we prove that under some assumptions

$$\int_{\mathbb{R}^{3N}} g(X)\rho_{\mathrm{BO}}(\mathbf{X})\,d\mathbf{X} = \int_{\mathbb{R}^{3N}} g(X)\rho_{\mathrm{S}}(\mathbf{X})\,dX + \mathcal{O}\big(M^{-1+\delta}\big)\,, \tag{6.2.12}$$

for any $\delta > 0$ and $g \in \mathcal{C}^3(\mathbb{R}^{3N})$. 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


[1] I. Babuška, A. Miller, and M. Vogelius. Adaptive methods and error estimation for elliptic problems of structural mechanics. In *Adaptive computational methods for partial differential equations (College Park, Md., 1983)*, pages 57–73. SIAM, Philadelphia, PA, 1983.

[2] Christian Bayer, Håkon Hoel, Petr Plecháč, Anders Szepessy, and Raúl Tempone. How accurate is molecular dynamics? *arXiv:1104.0953*, 2011.

[3] V. Bentkus. On the dependence of the Berry-Esseen bound on dimension. *J. Statist. Plann. Inference*, 113(2):385–402, 2003.

[4] Andrew C. Berry. The accuracy of the Gaussian approximation to the sum of independent variates. *Trans. Am. Math. Soc.*, 49:122–136, 1941.

[5] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637, 1973.

[6] William L. Briggs, Van Emden Henson, and Steve F. McCormick. *A multigrid tutorial. 2nd ed.* Philadelphia, PA: SIAM, Society for Industrial and Applied Mathematics., 2000.

[7] Russel E. Caflisch. Monte Carlo and quasi-Monte Carlo methods. *Acta Numerica*, 7(-1):1–49, 1998.

[8] Jesper Carlsson, Kyoung-Sook Moon, Anders Szepessy, Raúl Tempone, and Georgios Zouraris. Stochastic differential equations: Models and numerics. Lecture notes used at KTH, 2010.

[9] Y. S. Chow and Herbert Robbins. On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *Ann. Math. Statist.*, 36:457–462, 1965.

[10] R. H. Clarke. A statistical theory of mobile-radio reception. *Bell Sys. Tech.*, 47:957–1000, 1968.

[11] Gregory Durgin. *Space-time wireless channels.* Prentice Hall Press, Upper Saddle River, NJ, USA, 2002.

[12] Richard Durrett. *Probability: theory and examples.* Duxbury Press, Belmont, CA, second edition, 1996.

[13] Kenneth Eriksson, Don Estep, Peter Hansbo, and Claes Johnson. Introduction to adaptive methods for differential equations. In *Acta numerica, 1995*, Acta Numer., pages 105–158. Cambridge Univ. Press, Cambridge, 1995.

[14] Carl-Gustav Esseen. Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law. *Acta Math.*, 77:1–125, 1945.

[15] L. D. Faddeev and O. A. Yakubovskiĭ. *Lectures on quantum mechanics for mathematics students. Translated by Harold McFaden.* Student Mathematical Library 47. Providence, RI: American Mathematical Society (AMS). xii, 2009.

[16] Michael B. Giles. Multilevel Monte Carlo path simulation. *Oper. Res.*, 56(3):607–617, 2008.

[17] Paul Glasserman. *Monte Carlo methods in financial engineering*, volume 53 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 2004. Stochastic Modelling and Applied Probability.

[18] Stephen J. Gustafson and Israel Michael Sigal. *Mathematical concepts of quantum mechanics. 2nd ed.* Universitext. Berlin: Springer. xiii, 2011.

[19] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration. Structure-preserving algorithms for ordinary differential equations. Reprint of the second 2006 ed.* Springer Series in Computational Mathematics, 31. Berlin: Springer. xviii, 2010.

[20] Håkon Hoel, Erik Schwerin, Anders Szepessy, and Raúl Tempone. Adaptive multilevel monte carlo simulation. In *Numerical Analysis of Multiscale Computations*, volume 82 of *Lecture Notes in Computational Science and Engineering*, pages 217–234. Springer Berlin Heidelberg, 2012.

[21] E. Jouini, J. Cvitanić, and Marek Musiela, editors. *Option pricing, interest rates and risk management*. Handbooks in Mathematical Finance. Cambridge University Press, Cambridge, 2001.

[22] Peter E. Kloeden and Eckhard Platen. *Numerical solution of stochastic differential equations*, volume 23 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1992.

[23] Elliott H. Lieb and Robert Seiringer. *The stability of matter in quantum mechanics.* Cambridge: Cambridge University Press. xv, 2009.

[24] Robert C. Merton. Theory of rational option pricing. *The Bell Journal of Economics and Management Science*, 4(1):141–183, 1973.

[25] Kyoung-Sook Moon, Anders Szepessy, Raúl Tempone, and Georgios E. Zouraris. Convergence rates for adaptive weak approximation of stochastic differential equations. *Stoch. Anal. Appl.*, 23(3):511–558, 2005.

[26] Kyoung-Sook Moon, Erik von Schwerin, Anders Szepessy, and Raúl Tempone. An adaptive algorithm for ordinary, stochastic and partial differential equations. In *Recent advances in adaptive computation*, volume 383 of *Contemp. Math.*, pages 325–343. Amer. Math. Soc., Providence, RI, 2005.

[27] John C. Neu. *Training manual on transport and fluids.* Providence, RI: American Mathematical Society (AMS), 2010.

[28] Bernt Øksendal. *Stochastic differential equations*. Universitext. Springer-Verlag, Berlin, sixth edition, 2003. An introduction with applications.

[29] Mattias Sandberg and Anders Szepessy. Convergence rates of symplectic Pontryagin approximations in optimal control theory. *M2AN Math. Model. Numer. Anal.*, 40(1):149–173, 2006.

[30] Anders Szepessy, Raúl Tempone, and Georgios E. Zouraris. Adaptive weak approximation of stochastic differential equations. *Comm. Pure Appl. Math.*, 54(10):1169–1214, 2001.

[31] Denis Talay and Luciano Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Anal. Appl.*, 8(4):483–509 (1991), 1990.

[32] David Tse and Pramod Viswanath. *Fundamentals of wireless communication*. Cambridge University Press, New York, NY, USA, 2005.

[33] Wikipedia. Nyquist-shannon sampling theorem. *http://en.wikipedia.org/wiki/Nyquist-Shannon_sampling_theorem*, 2012 (accessed April 10, 2012).

# Part II

# Included Papers

# Paper I

# GAUSSIAN COARSE GRAINING OF A MASTER EQUATION EXTENSION OF CLARKE'S MODEL

HÅKON HOEL AND HENRIK NYBERG

Abstract. We study the error and computational cost of generating output signal realizations for the channel model of a moving receiver in a scattering environment, as in Clarke's model, with the extension that scatterers randomly flip on and off. At micro scale, the channel is modeled by a Multipath Fading Channel (MFC) model, and by coarse graining the micro scale model we derive a macro scale Gaussian process model. Four algorithms are presented for generating stochastic signal realizations, one for the MFC model and three for the Gaussian process model. A computational cost comparison of the presented algorithms indicates that Gaussian process algorithms generate signal realizations more efficiently than the MFC algorithm does. Numerical examples of generating signal realizations in time independent and time dependent scattering environments are given, and the problem of estimating model parameters from real life signal measurements is also studied.

## 1. Introduction

We consider the Multipath Fading Channel (MFC) model with a transmitter fixed and the receiver moving with a constant speed in an urban environment with buildings obstructing the line of sight between scatterer and receiver. Incoming rays at the receiver are thus modeled as scattered off the receiver's surroundings. Looking at scenarios where the distance between transmitter and receiver is large and the majority of scattering surfaces are flat walls, the vertical angle of arrival of incoming rays at the receiver is assumed to be 0 degrees, cf. Figure 1, i.e. scatterers are assumed to lie in the horizontal plane. The receiver receives $M$ incoming signal rays whose horizontal angle of arrival $\{\alpha_m\}_{m=1}^M$ are distributed according to a prescribed scatterer density $p : [0, 2\pi) \to \mathbb{R}_+$, and the resulting baseband output signal is modeled by

$$Z_{t,M} = \frac{1}{\sqrt{M}} \sum_{m=1}^{M} a(\alpha_m, t) e^{-i(2\pi \, \mathrm{f_c} \, \tau(\alpha_m, t) + \theta_m(t))}. \tag{1}$$

Here $\mathrm{f_c}$ denotes the carrier frequency, $\tau$ the delay function, and $a(\alpha_m, t)$ and $\theta_m(t)$ are respectively amplitude and phase shift random processes further described in Subsection 1.1. Two different delay functions are used in this paper; the full delay function depending on an explicit description of the scattering boundary, and $\tau(\alpha, t) = -vt\cos(\alpha)/c$, with $v$ and $c$ denoting the speed of the moving receiver and the speed light, respectively. The latter delay function is a first order approximation of the full delay function, cf. Appendix B.

2000 *Mathematics Subject Classification.* Primary 94A12; Secondary 60G15.

*Key words and phrases.* Wireless channel modeling; signal theory; master equations; Gaussian processes.

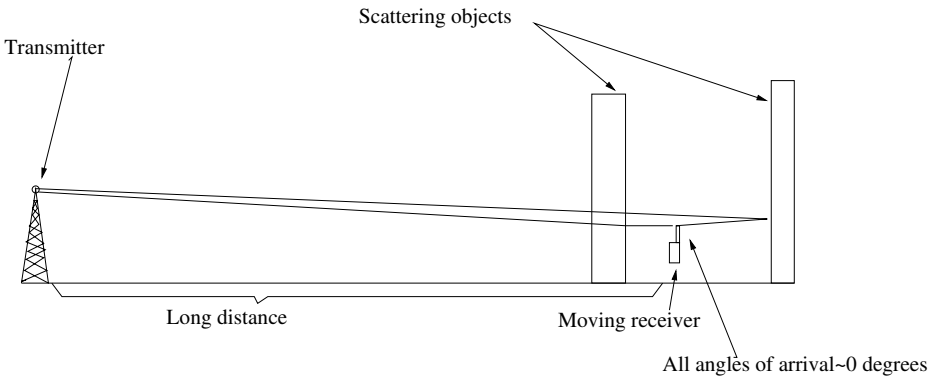


FIGURE 1. Illustration of the typical scattering environment which we wish to model the wireless channel in.

The choice $\tau(\alpha, t) = -vt\cos(\alpha)/c$, is common in channel modeling since it results in a Wide Sense Stationary (WSS) channel model, i.e. a model whose signal realizations $Z_{t,M}$ have constant expected value as a function of time and the autocorrelation function $E[Z_{t+\Delta t,M} Z^*_{t,M}]$ is a function of only $\Delta t$, and therefore are, relatively speaking, easier to analyze.

To put the output of (1) in context, a more general input/output baseband signal representation is given by

$$Z_t = \int_{\mathbb{R}} X(t-\tau)h(t,\tau)d\tau,$$

where $X(t)$ denotes the baseband input signal and $h(t,\tau)$ the impulse response function. For the output in (1), the baseband input is $X := 1$, which corresponds to the zero bandwidth passband input signal $X(t) = \exp(-i2\pi\,\mathrm{f_c}\,t)$ demodulated to a 0 frequency signal, and the impulse response function is given by

$$h(t,\tau) = \frac{1}{\sqrt{M}} \sum_{m=1}^{M} a(\alpha_m, t) e^{-i(2\pi\,\mathrm{f_c}\,\tau(\alpha_m,t)+\theta_m)} \delta(t - \tau(\alpha_m, t))$$

with $\delta$ denoting the Dirac delta function. Note that although we model channels with single frequency input signal, Doppler effects deriving from the receiver moving relative to its surroundings will result in output signals having non-zero frequency bandwidth.

1.1. **The amplitude random process.** Local shadowing by reflecting objects in motion, for example cars, pedestrians and shaking leaves, causes scatterers to flip from being active to passive and vice versa. Attempting to include local shadowing in our MFC model, we define the amplitude random flip process $a(\alpha, t)$ which flips on when it changes value from 0 to $a^+(\alpha, t) \geq 0$ and off when the opposite change occurs. It is here assumed that the mapping $a^+ : [0, 2\pi) \times \mathbb{R} \to \mathbb{R}_+$, which represents the active state of a reflector, is piecewise continuous (it may for example be piecewise constant or depend on the distance from scatterer to receiver). The flip process (of a scatterer) is modeled as Poisson process with constant flip rate $C$:

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

where flips are independent from the phase shift processes $\{\theta_m(t)\}_{m=1}^M$ and from the scatterers' state $\bigotimes_{m=1}^M\{0, a^+(\alpha_m, t)\}$. The scatterers' initial state $\{a(\alpha_m, 0)\}_{m=1}^M$ is sampled according to the i.i.d. Bernoulli distribution $P(a(\alpha_m, 0) = a^+(\alpha, 0)) = 1/2$ and $P(a(\alpha_m, 0) = 0) = 1/2$ which is consistent with the steady state distribution as $t \to \infty$. For an illustration of the effect of a positive flip rate on an output signal realization and a comparison to a real life signal measurement, see Figure 2.

The phase shift processes $\{\theta_m(t)\}_{m=1}^M$ are at all times i.i.d. uniformly distributed in $[0, 2\pi)$. This is motivated from the assumption that wave lengths are very small compared to the wave paths' travel distance from transmitter to receiver. The phase of $\theta_m(t)$ is updated by sampling $\theta_m(t) \sim U(0, 2\pi)$ every time scatterer $m$ flips, since we assume that a scatterer flipping at a given angle $\alpha_m$ is equivalent to a new scatterer appearing at the given angle; a new scatterer requiring a new i.i.d. phase shift uniformly distributed in $[0, 2\pi)$.

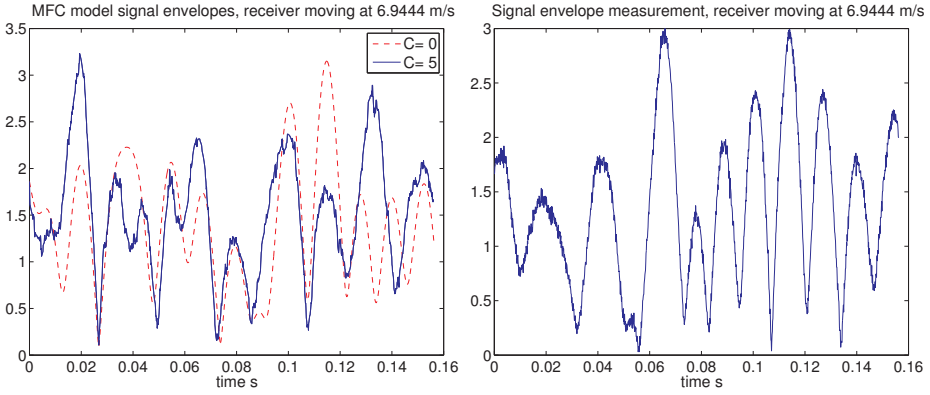


FIGURE 2. **Left plot:** Signal envelopes of two computer generated signal realizations using Algorithm 1 with the same random seed initialization. Both realizations are generated with $f_c = 1.8775$GHz, $v = 6.944$m/s, $a^+ = 2$, $p = (2\pi)^{-1}$ and $\tau(\alpha, t) = -vt\cos(\alpha)/c$. The only varying model parameter is the flip rate with $C = 0$ (dashed line) and $C = 5$ (full line). **Right plot:** Measured urban environment signal envelope whose carrier frequency and receiver speed are identical to the corresponding values for the realizations in the left plot.

**Remark 1.1.** *The presence of measurement noise in the data might be difficult to distinguish from flipping scatterer noise and can effect modeling parameter estimates, such as the flip rate. In our model, we assume that measurement noise is negligible relative to the noise generated by flipping scatterers.*

1.2. **Motivation for Gaussian processes model.** The MFC model with flips introduced here can be linked to the Master Equation, which is an equation often used in chemistry to describe the evolution of state space probabilities, cf. [13]. In a Master Equation setting, one typically assigns finite state spin variables on a lattice with probabilistic spin interaction dynamics and describe the time evolution of the probability to occupy each one of a discrete set of states through a differential

equation called the Master Equation. In our setting, we have the lattice $\{\alpha_m\}_{m=1}^M$ and the possibly time dependent state space $\bigotimes_{m=1}^M \{0, a^+(\alpha_m, t)\}$.

Creating discrete signal realizations of (1) straightforwardly by the MFC algorithm described in Section 2 is computationally very costly. This motivated us to try to construct an algorithm which reduced the computational cost while at the same time preserved the desired signal properties by using a new result in the Master Equation setting: In [11], Katsoulakis and Szepessy developed theory for coarse graining two state spin variables on a micro scale lattice into a macro scale Stochastic Differential Equation (SDE) representation which reduces the computational cost for such problems considerably. Using their theory, we developed a similar transition from the MFC model with flip state space on a micro scale lattice to a coarse grained SDE supposed to represent the output signal. However, comparing signal realizations of the Master Equation developed SDE to fine realizations of the MFC algorithm it was clear that the phase shift processes $\{\theta_m(t)\}_{m=1}^M$ were not resolved in the SDE model. This observation made us believe that a Master Equation SDE is not suitable coarse graining of (1) and lead us to instead try coarse graining with Gaussian processes, which are more general stochastic processes than SDEs.

The first outcome of coarse graining with Gaussian processes is Theorem 3.4 which shows that as $M \to \infty$, the signal $Z_{t,M}$ converges in distribution to a Gaussian process $Z_t$. Based on this theorem, we develop three algorithms, Algorithm 2, 3 and 4, using covariance and spectral properties to generate realizations of the limit Gaussian process. The developed algorithms are studied in terms of accuracy and computational cost, and a summary of this study, presented in Section 5, indicate that that the Gaussian process algorithms generate signal realizations more efficiently than the MFC algorithm, Algorithm 1, does.

1.3. **Related works and historical remarks.** In 1968, Clarke introduced the MFC model now known as Clarke's model in his seminal paper [7]. He considered the superposition of $M$ incoming waves

$$\xi_{t,M} = \frac{1}{\sqrt{M}} \sum_{m=1}^{M} e^{-i(2\pi \, \mathrm{f_c} \, v \cos(\alpha_m) t/c + \theta_m)}, \tag{2}$$

where, as before, $\mathrm{f_c}$ is the carrier frequency, $\{\theta_m\}_{m=1}^M$ are i.i.d. initial phase shifts and $\alpha_m$ is the arrival angle of the $m^{th}$ component wave distributed according to a given azimuth density $p(\alpha)$. Note in particular that this is a WSS model with delay function $\tau(\alpha_m, t) = -v \cos(\alpha_m) t/c$ which is identical to the delay function we mainly use in our MFC model. Considering the scenario with angle density $p(\alpha) = (2\pi)^{-1}$, Clarke noted that the auto correlation function, $E[\xi_{t,M} \xi^*_{0,M}]$, converges to the zeroth-order Bessel function of the first kind, $J_0(2\pi \, \mathrm{f_c} \, vt/c)$, as $M \to \infty$. And, further, he showed that its Power Spectral Density (PSD), which describes the frequency spread around the carrier frequency and is defined by the Fourier transform of the autocorrelation function, is on the form

$$S(f) = \begin{cases} \frac{c}{\pi \sqrt{(v \, \mathrm{f_c})^2 - (cf)^2}} & |f| < v \, \mathrm{f_c} \, / c \\ 0 & |f| \geq v \, \mathrm{f_c} \, / c. \end{cases} \tag{3}$$

This particular PSD is often referred to as Jakes' spectrum. For plots of these results, see Figure 3.

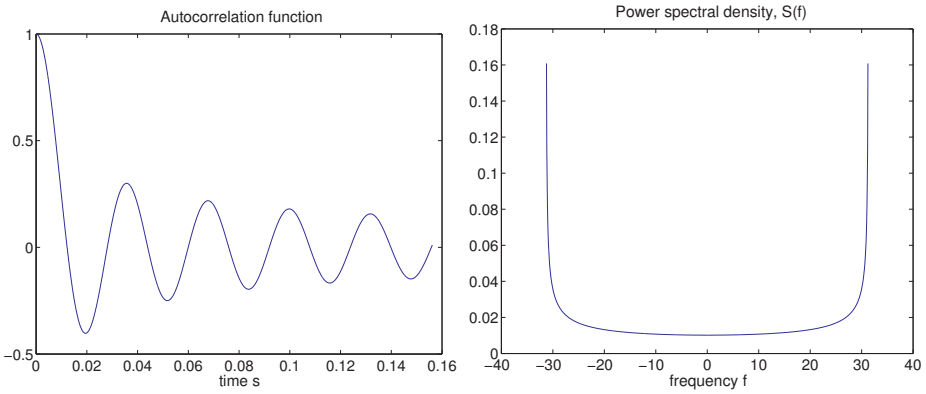


FIGURE 3. **Left plot:** The autocorrelation function for Clarke's model with azimuth density $p(\alpha) = (2\pi)^{-1}$, $v = 5\text{m/s}$ and $\mathrm{f_c} = 1.8775\text{GHz}$. **Right plot:** The power spectral density of Clarke's model, often called Jakes' spectrum, with the same model parameters as for the left plot.

Among the papers linking the MFC model to SDE which motivated this work, Feng, Field and Haykin [10] studied the output signal

$$\xi_{t,M} = \sum_{m=1}^{M} a_m e^{i\phi_m(t)}. \tag{4}$$

Here $a_m = O(M^{-1/2})$ are constant amplitudes and $\phi_m$ are phases with uniform random initialization on the interval $[0, 2\pi)$, and, with $dW_{t,m}$ denoting the Wiener increment and $B$ a positive constant, they assumed the phases also satisfied the increment relation

$$d\phi_m(t) = \sqrt{B} dW_{t,m}. \tag{5}$$

Since this is a model for stationary wireless channels, the phases do not include Doppler shift terms. As $M \to \infty$, they derived that $\xi_{t,M}$ converges to a complex valued Ornstein-Uhlenbeck process.

In [9], Feng and Field consider the MFC model with output

$$\xi_{t,M} = \sum_{m=1}^{M} a_m \exp\left(i(2\pi f_m t + \phi_t^{(m)})\right).$$

where the amplitudes $a_m$ are i.i.d. random variables and $f_m = \mathrm{f_D}\cos(\alpha_m)$ are Doppler shifts with $\mathrm{f_D}$ denoting the maximum Doppler shift and $\alpha_m$ the angle. The phases $\phi_t^{(m)}$ are independent Wiener processes with uniform initial distribution in $[0, 2\pi)$:

$$d\phi_t^{(m)} = \sqrt{B} dW_t^{(m)}, \qquad \phi_0^{(m)} \sim U[0, 2\pi),$$

where $B$ is a constant with the dimension of frequency. With this model they obtain the auto-correlation function

$$E[\varepsilon_{t,M}\varepsilon_{0,M}^*] = \sum_{m=1}^{M} E[a_m^2] e^{-B|t|/2} J_0(2\pi\,\mathrm{f_D}\,t),$$

and a corresponding PSD which they claim resemble measurements more than Clarke's model PSD. Although our modeling assumptions are quite different than Feng and Field's, we obtain similar autocorrelation and PSD results with our Gaussian Process model, cf. Section 4.

The development of flipping scatterers in our model is linked to the application of Poisson counting process as birth-death modeling of wave paths. Among the papers considering such models is Charalambous et al. [5] which study MFC models with output

$$\xi(t) = \sum_{m=1}^{M(T_s)} a_m(\tau_m) e^{i\phi_m(\tau_m,t)} x(t-\tau_m),$$

where $x(t)$ is the input signal, and $\tau_m, a_m(\tau_m)$ and $\phi_m(\tau_m, t)$ denote the propagation time delay, amplitude, and phase, respectively. The number of received paths, $M(T_s)$, is a Poisson counting process with $T_s > 0$ being a fixed stopping time. Under certain conditions, they obtain explicit expressions for the autocorrelation, PSD and moment-generating functions. In [16], Zwick et al. develop a temporally dynamic indoor channel model where birth and deaths of active wave paths are used to model the varying scattering environment, and in [6], Chong et al. introduce an indoor channel model where births and deaths of wave paths are generated by a Markov transition matrix.

All papers mentioned above present channel models with relatively few parameters, but there are also reports which have a highly physical approach thereby needing many parameters. The industry standard modeling report by 3GPP [1], considers reflectors gathered in clusters and models the received signal as a superposition of wave paths hitting the reflectors. Among its input parameters are: transmitter antenna and receiver antenna orientation, line of sight angle of departure from transmitter to receiver, angle of departure for every path from transmitter, angle of arrival for every path at receiver, transmitter velocity, angle of transmitter velocity, etc.

The rest of this paper is organized as follows. In Section 2, we present a numerical algorithm for generating signal realizations of our proposed MFC model. Section 3 motivates theoretically approximating the output signal by a Gaussian process model and two algorithms for generating Gaussian process signal realizations based on the covariance matrix is developed. An error and complexity analysis of Algorithm 1, 2 and 3 is also included. In Section 4, we investigate the relation between the signal's autocorrelation and PSD for WSS signals. A method for estimating the flip rate and scatterer density from PSD measurements is described, and a PSD based algorithm for generating Gaussian process signal realizations is presented. Section 5 summarizes the complexity analysis results obtained for the five algorithms considered in this paper. Section 6 concludes the paper with a numerical examples of signal realizations for time independent and time dependent scattering environments.

## 2. The MFC algorithm

We first present a numerical algorithm for generating signal realizations by the MFC model equation (1) on a set of sampling times $\mathbf{t} = (t_1, t_2, \ldots, t_N)$.

**Remark 2.1.** *It is possible to extend the MFC model and algorithm to scenarios including line of sight wave components, often called specular rays. Suppose you*

**Algorithm 1** The MFC algorithm

**Input:** Amplitude function $a^+$, flip rate $C$, carrier frequency $\mathrm{f_c}$, scatterer density $p$, receiver speed $v$, sampling times $\mathbf{t} = (t_1, t_2, \ldots, t_N)$.
**Output:** Signal realization $Z_{\mathbf{t},M} = (Z_{t_1,M}, Z_{t_2,M}, \ldots, Z_{t_N,M})$.
Generate a set of i.i.d angles of arrival $\{\alpha_k\}_{k=1}^M$ distributed according to the density $p(\alpha)$.
Generate a set of i.i.d. phase shifts $\{\theta_k(0)\}_{k=1}^M$ with $\theta_k(0) \sim U(0, 2\pi)$.
Generate the initial state of the amplitudes $\{a(\alpha_k, t_1)\}_{k=1}^M$ which are i.i.d. restricted to the steady state initial condition $P(a(\alpha, t_1) = 0) = P(a(\alpha, t_1) = a^+(\alpha, t_1)) = 1/2$.
Compute $Z_{t_1,M}$ according to (1).
**for** $j = 2$ to $N$ **do**
  **for** $k = 1$ to $M$ **do**
    Generate $n_k \sim Poisson(C(t_j - t_{j-1}))$.
    Flip the value of $a(\alpha_k, t_j)$ $n_k$ times.
    If $n_k > 0$, update the phase shift process by generating a new random phase shift $\theta_k(t_j) \sim U(0, 2\pi)$.
  **end for**
  Compute $Z_{t_j,M}$ according to (1).
**end for**

*have the input/output relation consisting of many diffuse ray contributions in $Z_{t,M}$ and one specular ray incoming from the angle 0 with amplitude $V$. Then an output signal can be generated by*

$$Z_{t,M} + Ve^{-i(2\pi\,\mathrm{f_c}\,\tau(0,t))},$$

*where $Ve^{-i(2\pi\,\mathrm{f_c}\,\tau(\alpha,t))}$ is a deterministic term modeling the specular ray contribution.*

## 3. Stochastic model for the signal with static scatterers

In this section we will show that the normalized signal $Z_{t,M}$ of equation 1 converges in distribution to a complex Gaussian process as $M \to \infty$. Thereafter, the covariance of the limit Gaussian process is derived and used to construct an algorithm for generating signal realizations in Algorithm 2. But first, let us recall the definitions of multivariate complex normal distributions and Gaussian processes.

**Definition 3.1** (Multivariate complex normal distribution I)**.** *Suppose $X$ and $Y$ are random vectors in $\mathbb{R}^n$ such that $(X, Y)$ is a $2n-$dimensional normal random vector. Then we say that $Z = X + iY$ is complex normal distributed and its distribution is described by the mean $\mu = E[Z]$, the covariance matrix $K = E[(Z-\mu)(Z-\mu)^H]$ and the pseudo-covariance matrix $J = E[(Z-\mu)(Z-\mu)^T]$, where $^T$ denotes the transpose operator and $^H$ the Hermitian operator. We write $Z \sim \mathcal{N}_{\mathbb{C}}(\mu, K, J)$.*

An alternative definition for multivariate complex normal distributions, which often is easier to work with, derives from the one-to-one relation between the characteristic function and the distribution:

**Definition 3.2** (Multivariate complex normal distribution II)**.** *A random vector $Z = (Z_1, Z_2, \ldots, Z_n) \in \mathbb{C}^n$ is said to be complex multivariate normal distributed if*

*the linear combination of its components,* $\mathbf{c}^H Z \in \mathbb{C}^1$ *is complex normal distributed for all* $\mathbf{c} \in \mathbb{C}^n$.

**Definition 3.3** (Complex Gaussian process)**.** *A complex Gaussian process is a stochastic process* $\{Z_t\}_{t\in[0,T)}$, $Z_t \in \mathbb{C}^1$, *for which any finite length sample vector* $(Z_{t_1}, Z_{t_2}, \ldots, Z_{t_n})$ *with* $0 \leq t_1 \leq t_2 \leq \ldots \leq t_n < T$ *is complex normal distributed.*

With the aid of the Central Limit Theorem (CLT), we now show that $Z_{t,M}$ converges in distribution to a Gaussian process as $M \to \infty$.

**Theorem 3.4** (Model (1)'s distributional convergence to a Gaussian process)**.** *Assume the amplitude function* $a^+(\cdot, t)$ *and the delay function* $\tau(\cdot, t)$ *is bounded and piecewise continuous on* $[0, 2\pi)$ *for all times* $t \in [0, T)$. *Then the signal*

$$Z_{t,M} := \frac{1}{\sqrt{M}} \sum_{m=1}^{M} a(\alpha_m, t) e^{-i(2\pi\, \mathrm{f_c}\, \tau(\alpha_m,t)+\theta_m(t))}$$

*converges in distribution to a complex Gaussian process* $Z_t$ *as* $M \to \infty$.

*Proof.* Letting $^*$ denote the complex conjugate, definitions 3.2 and 3.3 imply that if sums of the kind

$$\Upsilon_{\mathbf{t},\mathbf{c},M} := \sum_{i=1}^{n} c_i^* Z_{t_i,M}, \tag{6}$$

converge in distribution to a complex normal for all finite length sampling times $\mathbf{t} = (t_1, t_2, \ldots, t_n) \subset [0, T)$ and complex valued vectors $\mathbf{c} = (c_1, c_2, \ldots, c_n)$, then $Z_{t,M}$ converges in distribution to a complex Gaussian process on $[0, T)$.

Writing out equation (6), we have

$$\Upsilon_{\mathbf{t},\mathbf{c},M} = \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \underbrace{\sum_{j=1}^{n} c_j^* a(\alpha_m, t_j) e^{-i(2\pi\, \mathrm{f_c}\, \tau(\alpha_m,t_j)+\theta_m(t_j))}}_{=:\xi_m}.$$

Since both $\{\theta_m(\cdot)\}_m$, $\{a(\alpha_m, \cdot)\}_m$ are i.i.d. and mutually independent, the r.v. $\{\xi_m\}_m$ are also i.i.d. with mean

$$\mu = E[\xi_m] = \sum_{j=1}^{n} c_j^* E\left[ E[a(\alpha_m, t_j)|\alpha_m] \underbrace{E\left[ e^{-i(2\pi\, \mathrm{f_c}\, \tau(\alpha_m,t_j)+\theta_m(t_j))} \middle| \alpha_m \right]}_{=0} \right] = 0. \tag{7}$$

Before computing the covariance and pseudo-covariance of $\xi_m$, let us derive som useful properties. By the definition of the phase shift processes and amplitude process given in Subsection 1.1, we see that

$$P(\theta_m(t_j) = \theta_m(t_k)) = e^{-C|t_j - t_k|},$$

and

$$\begin{aligned} &E[a(\alpha_m, t_j)a(\alpha_m, t_k)|\alpha_m, \theta_m(t_j) = \theta_m(t_k)] \\ &\quad = a^+(\alpha_m, t_j)a^+(\alpha_m, t_k) P\left(a(\alpha_m, t_j) = a^+(\alpha_m, t_k)\right) \\ &\quad = \frac{a^+(\alpha_m, t_j)a^+(\alpha_m, t_k)}{2}. \end{aligned}$$

Consequently,

(8)
$$E\left[a(\alpha_m,t_j)a(\alpha_m,t_k)e^{i(2\pi\,\mathrm{f_c}(\tau(\alpha_m,t_k)-\tau(\alpha_m,t_j))+\theta_m(t_k)-\theta_m(t_j))}\Big|\alpha_m\right]$$
$$= P(\theta_m(t_j)=\theta_m(t_k))E[a(\alpha_m,t_j)a(\alpha_m,t_k)|\alpha_m,\theta_m(t_j)=\theta_m(t_k)]e^{i2\pi\,\mathrm{f_c}(\tau(\alpha_m,t_k)-\tau(\alpha_m,t_j))}$$
$$= e^{-C|t_j-t_k|}\frac{a^+(\alpha_m,t_j)a^+(\alpha_m,t_k)}{2}e^{i2\pi\,\mathrm{f_c}(\tau(\alpha_m,t_k)-\tau(\alpha_m,t_j))}.$$

and

$$E[a(\alpha_m,t_j)a(\alpha_m,t_k)e^{-i(2\pi\,\mathrm{f_c}(\tau(\alpha_m,t_k)+\tau(\alpha_m,t_j))+\theta_m(t_k)+\theta_m(t_j))}|\alpha_m]=0. \tag{9}$$

The covariance of $\xi_m$ is derived using (8)

(10)
$$K = E[|\xi_m|^2]$$
$$= \sum_{j,k=1}^{n} c_j^* c_k E\left[E\left[a(\alpha_m,t_j)a(\alpha_m,t_k)e^{i(2\pi\,\mathrm{f_c}(\tau(\alpha_m,t_k)-\tau(\alpha_m,t_j))+\theta_m(t_k)-\theta_m(t_j))}\Big|\alpha_m\right]\right]$$
$$= \sum_{j,k=1}^{n} c_j^* c_k \frac{e^{-C|t_j-t_k|}}{2}\int_0^{2\pi} a^+(\alpha,t_j)a^+(\alpha,t_k)e^{i2\pi\,\mathrm{f_c}(\tau(\alpha,t_k)-\tau(\alpha,t_j))}p(\alpha)\,d\alpha,$$

and the pseudo-covariance from using (9),

(11)
$$J = E[\xi_m^2]$$
$$= \sum_{j,k=1}^{n} c_j^* c_k E\left[E\left[a(\alpha_m,t_j)a(\alpha_m,t_k)e^{-i(2\pi\,\mathrm{f_c}(\tau(\alpha_m,t_k)+\tau(\alpha_m,t_j))+\theta_m(t_k)+\theta_m(t_j))}\Big|\alpha_m\right]\right]$$
$$= 0.$$

Having shown that $\Upsilon_{\mathbf{t},\mathbf{c},M} = \sum_{m=1}^{M}\xi_m/\sqrt{M}$ is a sum of i.i.d. complex valued random variables $\xi_m$ with mean $\mu = E[\xi_m] = 0$, pseudo-covariance $J$ and the bounded covariance $K$ as given in equation (11) and (10), respectively, it follows by the CLT for complex valued random variables that the normalized sums $\Upsilon_{\mathbf{t},\mathbf{c},M}$ converge in distribution to the complex normal $\mathcal{N}_{\mathbb{C}}(\mu,J,K)$ as $M\to\infty$. □

Having proved that $Z_{t,M}$ converges in distribution to a complex Gaussian process $Z_t$, our next goal is to construct an algorithm that generates realizations $Z_{\mathbf{t}} = (Z_{t_1}, Z_{t_2},\ldots,Z_{t_N})^T$ of the process sampled at a set of times $\mathbf{t} = (t_1,t_2,\ldots,t_N)$. To achieve that goal, we must first describe $Z_{\mathbf{t}}$ in terms of its mean, pseudo-covariance and covariance. Consider, as in the proof of Theorem 3.4, a sum $\Upsilon_{\mathbf{t},\mathbf{c}} = \lim_{M\to\infty}\sum_{j=1}^{N} c_j^* Z_{t_j,M}$ for a complex valued vector $\mathbf{c} = (c_1,c_2,\ldots,c_N)$. By choosing $\mathbf{c}$ such that $c_j = \delta_{jk}$, with $\delta_{jk}$ denoting the Kronecker delta, it follows from equation (7) and the proof of Theorem 3.4 that

$$\mu_k = E[Z_{t_k}] = E[\Upsilon_{\mathbf{t},\mathbf{c}}] = 0, \quad \text{forall} \quad k\in\{1,2,\ldots,N\}.$$

Choosing instead $c_j = c_k = 1$ and the other elements of $\mathbf{c}$ equal to 0, it follows from equations (11) and (10) that $J(t_j,t_k) = 0$ and

$$K(t_j,t_k) = \frac{e^{-C|t_j-t_k|}}{2}\int_0^{2\pi} a^+(\alpha,t_j)a^+(\alpha,t_k)e^{i2\pi\,\mathrm{f_c}(\tau(\alpha,t_k)-\tau(\alpha,t_j))}p(\alpha)\,d\alpha \tag{12}$$

for all $j, k \in \{1, 2, \ldots, N\}$.

When the pseudo-covariance $J = 0$, the multivariate complex normal is said to be circular symmetric and, for simplicity, its distribution is represented by $\mathcal{N}_{\mathbb{C}}(\mu, K)$. An $N-$dimensional circular symmetric complex normal $Z \sim \mathcal{N}_{\mathbb{C}}(\mu, K)$ has the density representation

$$P(Z = z) = \frac{e^{-(z-\mu)^H K^{-1}(z-\mu)}}{\det(K)\pi^N}.$$

For the sampled limit complex Gaussian process studied here, we thus see that $Z_{\mathbf{t}} \sim \mathcal{N}_{\mathbb{C}}(0, K)$ with the terms of $K$ given by (12).

3.1. **Gaussian process algorithm based on the covariance matrix.** We now present an algorithm that generates Gaussian process signal realizations on a set of $N$ sampling times $\mathbf{t}$ by multiplying the square root of the covariance matrix to a vector of i.i.d. standard complex normals, see (13). This is a standard way of generating Gaussian process realizations, c.f. [2], which requires having predetermined the covariance matrix $K$ integral terms given in (12). Generally, however, these covariance integral terms are not solvable by pen and paper, so we approximate them using numerical integration. This results in an approximation of the covariance matrix $K$, which we denote $\overline{\mathrm{K}}$. In the algorithm to be presented, Algorithm 2, $\overline{\mathrm{K}}$ is used to generate Gaussian process signal realizations $\overline{\mathrm{Z}}_{\mathbf{t}} = (\overline{\mathrm{Z}}_{t_1}, \overline{\mathrm{Z}}_{t_2}, \ldots, \overline{\mathrm{Z}}_{t_N})$.

**Algorithm 2** Covariance matrix based Gaussian process algorithm

**Input:** Amplitude function $a^+$, flip rate $C$, carrier frequency $\mathrm{f_c}$, scatterer density $p$, receiver speed $v$, sampling times $\mathbf{t} = (t_1, t_2, \ldots, t_N)$.
**Output:** Gaussian process realization $\overline{\mathrm{Z}}_{\mathbf{t}} = (\overline{\mathrm{Z}}_{t_1}, \overline{\mathrm{Z}}_{t_2}, \ldots, \overline{\mathrm{Z}}_{t_N})$.
**for** $i = 1$ to $N$ **do**
  **for** $j = 1$ to $N$ **do**
    Approximate the term $K(t_i, t_j)$ of equation (12) with quadrature and store the value in the approximate covariance matrix $\overline{\mathrm{K}}(t_i, t_j)$.
  **end for**
**end for**
Singular value decompose $\overline{\mathrm{K}} = \overline{U}\,\overline{S}\,\overline{U}^H$, and compute $\overline{\mathrm{K}}^{1/2} = \overline{U}\,\overline{S}^{1/2}$.
Generate a vector of $N$ i.i.d. standard complex normal elements; $\widehat{Z} \sim \mathcal{N}_{\mathbb{C}}(0, I_N)$, and thereafter the sampled output signal realization

$$\overline{\mathrm{Z}}_{\mathbf{t}} = \overline{\mathrm{K}}^{1/2}\,\widehat{Z}. \tag{13}$$

**Remark 3.5.** *As it were for the MFC algorithm, it is also possible to extend the covariance based Gaussian process algorithms to scenarios with specular ray contributions. Assume you have the input/output relation consisting of many diffuse ray contributions modeled by the Gaussian process output $\overline{\mathrm{Z}}_t$ and one specular ray incoming from the angle $\alpha$ with amplitude $V$ by*

$$\overline{\mathrm{Z}}_t + V e^{-i(2\pi\,\mathrm{f_c}\,\tau(\alpha,t))}.$$

3.2. **Error estimates and computational complexity.** Our next objective is to estimate the computational cost of generating signal realizations on a time grid $\mathbf{t} = (t_1, t_2, \ldots, t_N) \subset [0, T)$ satisfying an accuracy requirement when using either Algorithm 1 or 2. We first introduce a measure for how well, in distributional sense, an algorithm's stochastic realizations approximates the sampled limit Gaussian process $Z_{\mathbf{t}}$.

**Definition 3.6** (Distributional signal error measure)**.** *Let $K$ denote the covariance matrix of the sampled Gaussian process $Z_{\mathbf{t}} = (Z_{t_1}, Z_{t_2}, \ldots, Z_{t_N})$ expressed in equation* (12) *and let $\overline{\mathrm{Z}}_{\mathbf{t}}$ denote a generated signal realization that approximates $Z_{\mathbf{t}}$ in distributional sense. Then, under the assumption that $K$ is non-singular, the distributional error of $\overline{\mathrm{Z}}_{\mathbf{t}}$ is defined by*

$$e(\overline{\mathrm{Z}}_{\mathbf{t}}) := \sup_{A \in \mathfrak{C}(\mathbb{C}^N)} \left| P\left(\overline{\mathrm{Z}}_{\mathbf{t}} \in A\right) - P\left(Z_{\mathbf{t}} \in A\right) \right|,$$

*where $\mathfrak{C}(\mathbb{C}^N)$ denotes the class of convex sets in $\mathbb{C}^N$.*

3.3. **Computational cost of Algorithm 1.** To estimate the error for signals generated by Algorithm 1, we first review a theorem on the Berry-Essen bound's dimensional dependency.

**Theorem 3.7** (Bentkus [3])**.** *Let $\widehat{X}_i$ be i.i.d. random vectors in $\mathbb{R}^N$. Assume $\widehat{X}_i$ has mean zero and identity covariance matrix. Write $\beta = E[|\widehat{X}_i|^3]$, let*

$$S_M = \frac{1}{M} \sum_{m=1}^{M} \widehat{X}_m,$$

*and denote by $\mathfrak{C}(\mathbb{R}^N)$ the class of all convex sets in $\mathbb{R}^N$. Then*

$$\sup_{A \in \mathfrak{C}(\mathbb{R}^N)} \left| P\left(S_M \in A\right) - \int_A \frac{e^{-|x|^2/2}}{(2\pi)^{N/2}} dx \right| \leq \frac{400 N^{1/4} E[|\widehat{X}|^3]}{\sqrt{M}}.$$

**Remark 3.8.** *Considering the class of balls in $\mathbb{R}^N$ instead of the class of convex sets, it is possible to reduce the upper bound of Theorem 3.7 by a factor $O(N^{-1/4})$, c.f.* [3]*. However, for the processes studied in this paper, it is of interest to include at least all ellipsoidally shaped sets.*

Using Theorem 3.7 we next derive an upper bound for $e(Z_{\mathbf{t},M})$ as a function both of the number of scatterers $M$ and the number of sampling times $N$.

**Corollary 3.9** (Error bound for Algorithm 1)**.** *Let $Z_{\mathbf{t},M} = (Z_{t_1,M}, Z_{t_2,M}, \ldots, Z_{t_N,M})$ denote a sampled output signal generated by Algorithm 1 with sample times $\mathbf{t}$ restricted to $[0, T)$. Assume further the covariance matrix $K$ of the sampled limit Gaussian process $Z_{\mathbf{t}}$, cf. equation* (12)*, is non-singular so that it may be represented by the singular value decomposition $K = USU^H$ with $S = \mathrm{diag}(s_i)$, $s_1 \geq s_2 \geq \ldots \geq s_N > 0$. Then*

$$e(Z_{\mathbf{t},M}) \leq \frac{2^{3/4} \cdot 400\, N^{7/4} \left\| E\left[(a^+(\alpha, \cdot))^3\right] \right\|_{L^\infty([0,T))}}{s_N^{3/2} M^{1/2}}.$$

*Proof.* Since $Z_{\mathbf{t}}$ is a circular symmetric complex normal, it holds for any set $A \in \mathfrak{C}(\mathbb{C}^N)$ that

$$P\left(Z_{\mathbf{t},M} \in A\right) - P\left(Z_{\mathbf{t}} \in A\right) = P\left((K/2)^{-1/2} Z_{\mathbf{t},M} \in A\right) - \int_{(K/2)^{-1/2} A} \frac{e^{-|z|^2/2}}{(2\pi)^N}\, dz,$$

where $(K/2)^{-1/2}$ being a linear operator implies that also $(K/2)^{-1/2}A \in \mathfrak{C}(\mathbb{C}^N)$. The linearly transformed signal $(K/2)^{-1/2}Z_{\mathbf{t},M}$ may be written

$$(K/2)^{-1/2}Z_{\mathbf{t},M} = \frac{1}{\sqrt{M}}\sum_{m=1}^{M}(K/2)^{-1/2}\breve{Z}_m$$

with i.i.d. complex valued random vectors

$$\breve{Z}_m := \begin{bmatrix} a(\alpha_m,t_1)\exp\left(-i(2\pi\,\mathrm{f_c}\,\tau(\alpha_m,t_1)+\theta_m(t_1))\right) \\ a(\alpha_m,t_2)\exp\left(-i(2\pi\,\mathrm{f_c}\,\tau(\alpha_m,t_2)+\theta_m(t_2))\right) \\ \vdots \\ a(\alpha_m,t_N)\exp\left(-i(2\pi\,\mathrm{f_c}\,\tau(\alpha_m,t_N)+\theta_m(t_N))\right) \end{bmatrix}.$$

Following the proof of Theorem 3.4, $\breve{Z}_m$ has zero mean, zero pseudo-covariance matrix, and covariance matrix $K$, identical to the covariance matrix of $Z_{\mathbf{t}}$. This implies that the random vectors $(K/2)^{-1/2}\breve{Z}_m$ has mean zero, zero pseudo-covariance matrix, and covariance matrix equal to $2I_N$, with $I_N$ denoting the $N-$dimensional identity matrix. Then, representing $\mathbb{C}^N$ by $\mathbb{R}^{2N}$, we note that $2N-$dimensional real valued random vector $(\mathrm{Re}((K/2)^{-1/2}\breve{Z}_m), \mathrm{Im}((K/2)^{-1/2}\breve{Z}_m))$ has mean 0 and identity covariance $I_{2N}$, and the proof is completed by bounding the norm of $E[|(\mathrm{Re}((K/2)^{-1/2}\breve{Z}_m), \mathrm{Im}((K/2)^{-1/2}\breve{Z}_m))|^3] = E[|(K/2)^{-1/2}\breve{Z}_m|^3]$, and applying Theorem 3.7. By Hölder's inequality,

$$\begin{aligned} E[|(K/2)^{-1/2}\breve{Z}_m|^3] &\leq \frac{2^{3/2}}{s_N^{3/2}}E\left[\left(\sum_{j=1}^{N}(a(\alpha,t_j))^2\right)^{3/2}\right] \\ &\leq \frac{2^{3/2}\sqrt{N}}{s_N^{3/2}}\sum_{j=1}^{N}E\left[(a(\alpha,t_j))^3\right] \\ &= 2^{1/2}\left(\frac{N}{s_N}\right)^{3/2}\left\|E\left[(a^+(\alpha,\cdot))^3\right]\right\|_{L^\infty([0,T))}. \end{aligned}$$

End of proof. □

It follows from the above corollary that to generate a realization fulfilling

$$e(Z_{\mathbf{t},M}) \leq \mathrm{TOL}, \tag{14}$$

requires $M = O(\mathrm{TOL}^{-2}\,N^{7/2}S_{2N}^{-3})$, which has the computational cost $O(MN) = O(\mathrm{TOL}^{-2}(N^{3/2}/S_{2N})^3)$. For statistical purposes it is often interesting to generate many realizations, so we also note that the cost generating $L$ realizations fulfilling (14) is

$$\mathrm{Cost(Algorithm\ 1)} = O\left(L\,\mathrm{TOL}^{-2}\left(\frac{N^{3/2}}{S_{2N}}\right)^3\right).$$

3.4. **Computational cost of Algorithm 2.** Algorithm 2 uses the covariance matrix $\overline{\mathrm{K}}$ to generate realizations $\overline{\mathrm{Z}}_{\mathbf{t}} = (\overline{\mathrm{Z}}_{t_1}, \overline{\mathrm{Z}}_{t_2}, \ldots, Z_{t_N})$. In distribution, these realizations differ slightly from the realizations $Z_{\mathbf{t}}$ that would be obtained if using $Z_{\mathbf{t}}$'s covariance matrix $K$ to generate signals instead of $\overline{\mathrm{K}}$. Here, we will bound

the error $e(\overline{\mathrm{Z}}_{\mathbf{t}})$ in terms of the difference between $K$ and $\overline{\mathrm{K}}$, a difference which is a consequence of approximating the integral terms $K(t_j, t_k)$ of (12) by quadrature:

$$\overline{\mathrm{K}}(t_j, t_k) = \frac{e^{-C|t_j - t_k|}}{2} \sum_{\ell=1}^{M} a^+(x_\ell, t_j) a^+(x_\ell, t_k) e^{i2\pi\, \mathrm{f_c}(\tau(x_\ell, t_k) - \tau(x_\ell, t_j))} p(x_\ell)\ \nu_\ell$$

for the quadrature points $0 = x_1 < x_2 < \ldots < x_M = 2\pi$ and $\nu_\ell$ being quadrature weights fulfilling $\sum_{\ell=1}^{M} \nu_\ell = 2\pi$. Using $M$ quadrature points, the error bound

$$\max_{1<j,k<N} |K_{j,k} - \overline{\mathrm{K}}_{j,k}\,| \leq \epsilon = O(M^{-\gamma}) \tag{15}$$

will be fulfilled, $\gamma > 0$ here depending on the quadrature method used. Consequently, we may write $\overline{\mathrm{K}} = K + \delta K$ with $\|\delta K\|_2 \leq N^{1/2}\epsilon$. Having described the difference between the covariance matrices we now state a theorem that bounds $e(\overline{\mathrm{Z}}_{\mathbf{t}})$ for circular symmetric complex normals $\overline{\mathrm{Z}}_{\mathbf{t}}$ approximating the sampled limit complex Gaussian process $Z_{\mathbf{t}}$.

**Theorem 3.10.** *Assume the covariance matrix $K$ of the sampled limit complex Gaussian process $Z_{\mathbf{t}} = (Z_{t_1}, Z_{t_2}, \ldots, Z_{t_N})$, given in (12), is non-singular so that it may be represented by the singular value decomposition $K = USU^H$ with $S = \mathrm{diag}(s_i)$, $s_1 \geq s_2 \geq \ldots \geq s_N > 0$. Let further $\overline{\mathrm{Z}}_{\mathbf{t}}$ be a circular symmetric normal Gaussian with mean 0 and covariance matrix $\overline{\mathrm{K}}$, and assume $\|K - \overline{\mathrm{K}}\,\|_2 \leq N^{1/2}\epsilon$, with $\epsilon = O(M^{-\gamma})$, $\gamma > 0$ the convergence order of the quadrature method, and $M$ chosen so large that $10N^{3/2}\epsilon < s_N$. Then*

$$e(\overline{\mathrm{Z}}_{\mathbf{t}}) = O\left(\frac{N^{3/2}}{s_N\, M^{\gamma}}\right).$$

*Proof.* By construction, $\overline{\mathrm{K}}$ is symmetric, so it has a singular value decomposition $\overline{\mathrm{K}} = \overline{U}\,\overline{S}\,\overline{U}^H$, with $\overline{s}_1 \geq \overline{s}_2 \geq \ldots \overline{s}_N \in \mathbb{R}$. According to Corollary A.1 and the assumption on $M$ being sufficiently large,

$$\max_{1\leq j\leq N} |s_j - \overline{s}_j| \leq N^{1/2}\epsilon < s_N.$$

This implies that $\overline{s}_N > 0$ so that $\overline{\mathrm{K}}$ also is non-singular with

$$\overline{\mathrm{K}}^{-1} = (K + \delta K)^{-1} = (I + K^{-1}\delta K)^{-1} K^{-1} = K^{-1} + \sum_{j=1}^{\infty} (-K^{-1}\delta K)^j K^{-1}.$$

For later reference we further note that by the assumption of $M$ being sufficiently large,

$$\left\| K^{1/2} \sum_{j=1}^{\infty} (-K^{-1}\delta K)^j K^{-1/2} \right\|_2 \leq \frac{2N^{1/2}\epsilon}{s_N}. \tag{16}$$

For all Borel sets $A \subset \mathbb{C}^N$, we derive the following upper bound

$$
\begin{aligned}
P\left(\overline{\mathrm{Z}}_{\mathbf{t}} \in A\right) - P\left(Z_{\mathbf{t}} \in A\right) &= \frac{1}{\pi^N |\det(K)|} \int_A \frac{e^{-z^H \overline{\mathrm{K}}^{-1} z}}{|\det(K^{-1}\overline{\mathrm{K}})|} - e^{-z^H K^{-1} z}\, dz \\
&= \frac{1}{\pi^N |\det(K)|^{1/2}} \int_A e^{-z^H K^{-1} z} \left( \frac{e^{-z^H \sum_{j=1}^{\infty} (-K^{-1}\delta K)^j K^{-1} z}}{|\det(K^{-1}\overline{\mathrm{K}})|} - 1 \right) dz \\
&= \frac{1}{(2\pi)^N} \int_{(K/2)^{-1/2} A} e^{-|z|^2/2} \left( \frac{e^{-z^H K^{1/2} \sum_{j=1}^{\infty} (-K^{-1}\delta K)^j K^{-1/2} z/2}}{|\det(K^{-1}\overline{\mathrm{K}})|} - 1 \right) dz \\
&\overbrace{\leq}^{(16)} \frac{1}{(2\pi)^N} \int_{(K/2)^{-1/2} A} e^{-|z|^2/2} \left( \frac{e^{(N^{1/2}\epsilon/s_N)|z|^2}}{\prod_{j=1}^N \left(1 - N^{1/2}\epsilon/s_j\right)} - 1 \right) dz \\
&\leq \frac{1}{(2\pi)^N} \int_{\mathbb{R}^{2N}} e^{-|x|^2/2} \left( \left(1 - \frac{N^{1/2}\epsilon}{s_N}\right)^{-N} e^{(N^{1/2}\epsilon/s_N)|x|^2} - 1 \right) dx \\
&= \left(1 - \frac{N^{1/2}\epsilon}{s_N}\right)^{-N} \left(1 - 2\frac{N^{1/2}\epsilon}{s_N}\right)^{-N} - 1 \\
&\leq e^{10 N^{3/2}\epsilon/s_N} - 1 \\
&\leq \frac{20 N^{3/2}\epsilon}{s_N}.
\end{aligned}
$$

Here we used if $\widehat{X}$ is a $2N-$dimensional real valued multivariate normal with mean zero and identity covariance, then $E\left[e^{t|\widehat{X}|^2}\right] = (1-2t)^{-N}$ is the moment generating function to a chi-square distributed variable with $2N$ degrees of freedom. Calculating along the same lines one obtains the lower bound

$$
\begin{aligned}
P\left(\overline{\mathrm{Z}}_{\mathbf{t}} \in A\right) - P\left(Z_{\mathbf{t}} \in A\right) &\geq \left(1 + \frac{N^{1/2}\epsilon}{s_N}\right)^{-N} E\left[e^{-(N^{1/2}\epsilon/s_N)|\widehat{X}|^2}\right] - 1 \\
&\geq -\frac{3N^{3/2}\epsilon}{s_N}.
\end{aligned}
$$

Since the upper and lower bound obtained are both valid for the class of all Borel sets $A \in \mathbb{C}^N$ and this class contains the class of convex sets $\mathfrak{C}(\mathbb{C}^N)$, the proof is finished. □

It follows from Theorem 3.10 that to fulfill the accuracy condition $e(\overline{\mathrm{Z}}_{\mathbf{t}}) \leq \mathrm{TOL}$ for realizations generated by Algorithm 2 requires

$$
M = O\left( \left( \frac{N^{3/2}}{s_N\ \mathrm{TOL}} \right)^{1/\gamma} \right).
$$

The cost of generating $L$ signal realizations fulfilling the above mentioned accuracy condition then becomes the sum of $O(MN^2)$ to compute the matrix elements of $\overline{\mathrm{K}}$ by quadrature, $O(N^3)$ to construct the square root of $\overline{\mathrm{K}}$ and and $O(LN^2)$ for creating $L$ signal realizations by the matrix vector multiplication $\overline{\mathrm{K}}^{1/2}\, \tilde{X}$:

$$
\mathrm{Cost(Algorithm\ 2)} = O\left( N^3 + N^2 \left( \frac{N^{3/2}}{s_N\ \mathrm{TOL}} \right)^{1/\gamma} + LN^2 \right).
$$

3.5. **A circulant-embedding based algorithm with error and computational cost analysis.** Computing the square root of $\overline{\mathrm{K}}$ in Algorithm 2 and generating the output signal by (13) were both computationally costly operations, and in this subsection we consider modeling settings where it is possible to improve the efficiency of these operations by circulant-embedding of the covariance matrix and application of finite Fourier methods. The idea presented here is adapted from the material in [2, Chapter XI].

In the modeling setting when

$$\tau(\alpha,t) = -v\cos(\alpha)t/c \quad \text{and} \quad a^+(\alpha,t) = a^+(\alpha), \tag{17}$$

the limit Gaussian process $Z_t$ is a WSS process which simplifies the structure of the covariance:

$$K(t_i,t_j) = E[Z_{t_i}Z^*_{t_j}] = \frac{e^{-C|t_i-t_j|}}{2}\int_0^{2\pi} a^+(\alpha)^2 e^{i2\pi\,\mathrm{f_D}\cos(\alpha)(t_i-t_j)}p(\alpha)d\alpha =: A(t_i-t_j), \tag{18}$$

where we recall that $\mathrm{f_D} = \mathrm{f_c}\,v/c$ is the maximum Doppler shift and $A(t)$ denotes the autocorrelation function. Sampling the limit Gaussian process $Z_t$ on a uniform time grid $\mathbf{t} = (t_1, t_2, \ldots, t_N)$, $t_{i+1} - t_i = \Delta t$, in setting (17), yields a circular symmetric complex normal $Z_{\mathbf{t}}$ with a covariance matrix $K$ that is Toeplitz. Using the short hand notation $A_j = A(j\Delta t)$, the covariance matrix has the structure

$$K = \begin{bmatrix} A_0 & A_1 & \cdot & A_{N-2} & A_{N-1} \\ A_{-1} & A_0 & \cdot & A_{N-3} & A_{N-2} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ A_{-(N-2)} & A_{-(N-3)} & \cdot & A_0 & A_1 \\ A_{-(N-1)} & A_{-(N-2)} & \cdot & A_{-1} & A_0 \end{bmatrix}.$$

To illustrate circulant-embedding of $K$, let us for simplicity assume $A_{-j} = A_j$, and embed $K$ as the upper left corner of a circulant of order $2N-2$:

$$\mathcal{C} = \begin{bmatrix} A_0 & A_1 & \cdot & A_{N-2} & A_{N-1} & A_{N-2} & A_{N-3} & \cdot & A_2 & A_1 \\ A_1 & A_0 & \cdot & A_{N-3} & A_{N-2} & A_{N-1} & A_{N-2} & \cdot & A_3 & A_2 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ A_{N-1} & A_{N-2} & \cdot & A_1 & A_0 & A_1 & A_2 & \cdot & A_{N-3} & A_{N-2} \\ A_{N-2} & A_{N-1} & \cdot & A_2 & A_1 & A_0 & A_1 & \cdot & A_{N-4} & A_{N-3} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ A_1 & A_2 & \cdot & A_{N-2} & A_{N-1} & A_{N-2} & A_{N-3} & \cdot & A_1 & A_0 \end{bmatrix}.$$

(In the general case when we only have $A_{-j} = A_j^*$, the circulant-embedding of $K$ will be of size $4N-4$.) The circulant matrix $\mathcal{C}$ has the eigendecomposition

$$\mathcal{C} = F\Lambda F^H, \tag{19}$$

where $F$ is the finite Fourier matrix of order $2N-2$ with elements $F_{jk} = e^{i2\pi(j-1)(k-1)/(2N-2)}/\sqrt{2N-2}$, and $\Lambda$ is the diagonal matrix of eigenvalues

$$\mathrm{diag}(\Lambda) = F\mathbf{a} \quad \text{with} \quad \mathbf{a} = (A_0, A_1, \ldots, A_{N-1}, A_{N-2}, \ldots, A_2, A_1)^T. \tag{20}$$

The representation (19) shows that provided all entries of $\Lambda$ are non-negative, we may write $\mathcal{C}^{1/2} = F\Lambda^{1/2}$. The signal generated by

$$Z_{\mathbf{t}} = RF\Lambda^{1/2}\widehat{Z},$$

with $\widehat{Z} \sim \mathcal{N}_{\mathbb{C}}(0, I_{2N-2})$ and $R : \mathbb{C}^{2N-2} \to \mathbb{C}^N$ defined by $Rz = (z_1, z_2, \ldots, z_N)^T$ for all $z \in \mathbb{C}^{2N-2}$, will have the sought covariance $K$. This leads us to the following algorithm for generating WSS Gaussian process signal realizations.

**Algorithm 3** Covariance based algorithm for WSS processes

**Input:** Flip rate $C$, maximum Doppler shift $\mathrm{f_D}$, scatterer density $p$, uniform grid of sampling times $\mathbf{t} = (t_1, t_2, \ldots, t_N)$.
**Output:** Gaussian process realization $\overline{\mathrm{Z}}_{\mathbf{t}} = (\overline{\mathrm{Z}}_{t_1}, \overline{\mathrm{Z}}_{t_2}, \ldots, \overline{\mathrm{Z}}_{t_N})$.
**for** $j = -(N-1)$ to $N-1$ **do**
Approximate the term $A_j$ by quadrature of equation (18) and store the approximation in $\overline{A}_j$.
**end for**
Determine $\overline{\Lambda}$ by computing

$$\mathrm{diag}(\overline{\Lambda}) = F(\overline{A}_0, \overline{A}_1, \ldots, \overline{A}_{N-1}, \overline{A}_{N-2}, \ldots, \overline{A}_2, \overline{A}_1)^T,$$

using the Fast Fourier Transform (FFT).
Generate a multivariate complex normal vector $\widehat{Z} \sim \mathcal{N}_{\mathbb{C}}(0, I_{2N-2})$ and thereafter use the FFT to compute the output realization

$$\overline{\mathrm{Z}}_{\mathbf{t}} = RF\overline{\Lambda}^{1/2}\widehat{Z}. \tag{21}$$

Since a sampled realization $\overline{\mathrm{Z}}_{\mathbf{t}}$ generated by Algorithm 3 is multivariate circular symmetric complex Gaussian with mean zero and covariance $\overline{\mathrm{K}}$, we may use Theorem 3.10 to conclude that, under the assumptions there stated, $e(\overline{\mathrm{Z}}_{\mathbf{t}}) = O\left(N^{3/2}/(s_N M^\gamma)\right)$. Alternatively, under the assumptions that $\Lambda$ is positive definite, $\|K - \overline{\mathrm{K}}\|_2 \leq N^{1/2}\epsilon$ with $\epsilon = O(M^\gamma)$ and $M$ chosen sufficiently large so that $N^{1/2}\,\mathrm{tr}(\Lambda^{-1})\epsilon < 1/2$, one may derive the second error bound

$$e(\overline{\mathrm{Z}}_{\mathbf{t}}) = O\left(\frac{N^{1/2}\,\mathrm{tr}(\Lambda^{-1})}{M^\gamma}\right)$$

from observing that for any Borel set $A \in \mathbb{C}^N$

$$\begin{aligned}
&\left|P\left(F\overline{\Lambda}^{1/2}\widehat{Z} \in A \times \mathbb{C}^{N-2}\right) - P\left(F\Lambda^{1/2}\widehat{Z} \in A \times \mathbb{C}^{N-2}\right)\right| \\
&= \left|P\left((\Lambda^{-1}\overline{\Lambda})^{1/2}\widehat{Z} \in \Lambda^{-1/2}F^H(A \times \mathbb{C}^{N-2})\right) - P\left(\widehat{Z} \in \Lambda^{-1/2}F^H(A \times \mathbb{C}^{N-2})\right)\right| \\
&= O\left(N^{1/2}\epsilon\,\mathrm{tr}(\Lambda^{-1})\right),
\end{aligned}$$

where $\mathrm{tr}(\Lambda^{-1}) := \sum_j \Lambda_j^{-1}$. The cost of generating $L$ signal realizations with Algorithm 3 fulfilling the accuracy $e(\overline{\mathrm{Z}}_{\mathbf{t}}) \leq \mathrm{TOL}$ becomes the sum of $O(MN)$ to compute the covariance elements $\{\overline{A}_j\}_{j=-N+1}^{N-1}$ of $\overline{\mathrm{K}}$ by quadrature, $O(N\log(N))$ to compute $\overline{\Lambda}$ and $O(LN\log(N))$ for creating $L$ signal realizations using FFT:

$$\mathrm{Cost(Algorithm\ 3)} = O\left\{N\left[\frac{N^{1/2}}{\mathrm{TOL}}\min\left(\frac{N}{s_N}, \mathrm{tr}(\Lambda^{-1})\right)\right]^{1/\gamma} + LN\log(N)\right\}.$$

For a tentative comparison of the magnitude of $N/s_N$ and $\mathrm{tr}(\Lambda^{-1})$, see Section 5.

## 4. Applications of the power spectral density

Being the Fourier dual of the covariance function, the PSD can, as the covariance, be used in algorithms to construct signal realizations and to study properties process properties. Here we derive the PSD for in the WSS process modeling setting

$$\tau(\alpha,t) = -v\cos(\alpha)t/c, \qquad a^+(\alpha,t) = a^+(\alpha), \quad \text{and} \quad C = O(1). \tag{22}$$

The shape of the PSD is linked to the flip rate and the scattering density and we will describe a method for estimating the flip rate $C$ from the PSD computed from real life signal measurements and also look into the relation between PSD and scattering density. At the end of this section, an algorithm for generating Gaussian process signal realizations using the PSD function is presented with cost and error analysis included.

We first derive an expression for the PSD, using ideas from [9] and [7]. The PSD is given by $S_C(f) = \mathcal{F}\{A(\cdot)\}$, where $A(t)$ is the autocorrelation of the limit complex Gaussian process $Z_t$, $\mathcal{F}(\cdot)$ denotes the Fourier Transform, and the subscript $C$ in $S_C$ emphasizes that the PSD depends on the flip rate. In the setting (22), it follows from (18) that the autocorrelation is on the form

$$A(t) = \frac{e^{-C|t|}}{2}\int_0^{2\pi} (a^+(\alpha))^2 e^{i2\pi\,\mathrm{f_D}\cos(\alpha)t}\,p(\alpha)\,d\alpha, \tag{23}$$

where we recall that $\mathrm{f_D} = vt/c$. By the Convolution theorem for Fourier transforms,

$$\begin{aligned} S_C(f) &= \mathcal{F}\{A(\cdot)\}(f) \\ &= \mathcal{F}\left\{\frac{e^{-C|\cdot|}}{2}\right\} * \mathcal{F}\left\{\int_0^{2\pi} (a^+(\alpha))^2 e^{i2\pi\,\mathrm{f_D}\cos(\alpha)\cdot}\,p(\alpha)\,d\alpha\right\}(f), \end{aligned}$$

where $*$ denotes the convolution operator. Observe next that

$$\mathcal{F}\{e^{-C|\cdot|}\}(f) = \frac{2C}{C^2 + (2\pi f)^2}, \quad \forall C > 0,$$

and

$$\begin{aligned} &\mathcal{F}\left\{\int_0^{2\pi} (a^+(\alpha))^2 e^{i2\pi\,\mathrm{f_D}\cos(\alpha)\cdot}\,p(\alpha)\,d\alpha\right\}(f) \\ &= \int_{\mathbb{R}}\int_0^{2\pi} (a^+(\alpha))^2 e^{i2\pi\,\mathrm{f_D}\cos(\alpha)t} p(\alpha)\,d\alpha\, e^{-i2\pi ft}\,dt \qquad (\alpha = -\cos^{-1}(s/\,\mathrm{f_D}) + \{0,\pi\}) \\ &= \int_{\mathbb{R}}\int_{-\mathrm{f_D}}^{\mathrm{f_D}} \frac{1}{\mathrm{f_D}\sqrt{1-(s/\,\mathrm{f_D})^2}}\Bigg(e^{i2\pi st}(a^+(-\cos^{-1}(s/\,\mathrm{f_D})))^2 p(-\cos^{-1}(s/\,\mathrm{f_D})) \\ &\qquad + e^{-i2\pi st}\Big(a^+\big(-\cos^{-1}(s/\,\mathrm{f_D})+\pi\big)\Big)^2 p\big(-\cos^{-1}(s/\,\mathrm{f_D})+\pi\big)\Bigg) e^{-i2\pi ft}\,dt\,ds \\ &= \frac{1_{|f|<\mathrm{f_D}}(f)}{\mathrm{f_D}\sqrt{1-(f/\,\mathrm{f_D})^2}}\Bigg((a^+(-\cos^{-1}(f/\,\mathrm{f_D})))^2 p(-\cos^{-1}(f/\,\mathrm{f_D})) \\ &\qquad + \Big(a^+\big(-\cos^{-1}(-f/\,\mathrm{f_D})+\pi\big)\Big)^2 p\big(-\cos^{-1}(-f/\,\mathrm{f_D})+\pi\big)\Bigg)\Bigg). \end{aligned}$$

Convolving the last two expressions leads to the following PSD integral expression

$$S_C(f) = \int_0^{2\pi} \frac{C(a^+(\alpha))^2 p(\alpha)}{C^2 + (2\pi(f - \mathrm{f_D}\cos(\alpha)))^2} d\alpha. \tag{24}$$

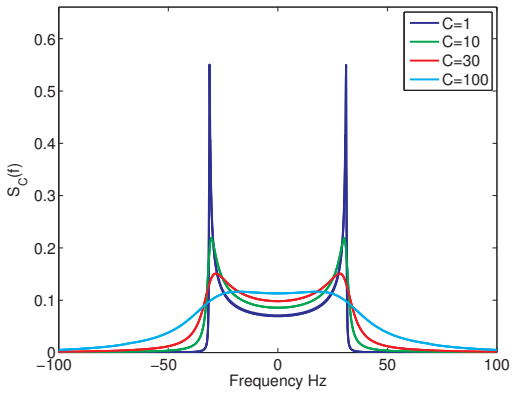


FIGURE 4. The PSD function $S_C(f)$ is plotted for different flip rate values $C$ with $\mathrm{f_D} = 50$Hz when $(a^+)^2 p = (2\pi)^{-1}$.

We see that $S_C(f)$ depends on the flip rate $C$ and the term $(a^+)^2 p$, which we hereafter refer to as a scaled scatterer density. Figure 4 illustrates $S_C(f)$'s dependency on the flip rate $C$ in the modeling settings $(a^+)^2 p = (2\pi)^{-1}$. In that setting $S_0(f)$ is Jakes' spectrum and $S_C(f)$ is a progressively mollified version of Jakes' spectrum the higher the value of $C$ is. In [9] it is remarked that the PSD $S_C(f)$ with $C = O(1)$ is more in accordance with real life measurements than Jakes' spectrum is.

4.1. **A link to Feng and Field's model.** Our MFC model's output signal autocorrelation and PSD results are very similar to what Feng and Field obtained under quite different modeling assumptions in [9]. They considered the modified Clarke's model

$$\varepsilon_t = \sum_{n=1}^{N} a_n \exp\left(i(2\pi f_n t + \phi_t^{(n)})\right),$$

where the amplitudes $a_n$ are i.i.d. random variables and $f_n = \mathrm{f_D}\cos(\alpha_n)$ are Doppler shifts where, as in this paper, $\mathrm{f_D} = \mathrm{f_c}\, v/c$. The phases $\phi_t^{(n)}$ are independent Wiener processes with uniform initial distribution in $[0, 2\pi)$:

$$d\phi_t^{(n)} = \sqrt{B} dW_t^{(n)}, \quad \phi_0^{(n)} \sim U[0, 2\pi), \quad \forall n \in \{1, 2, \ldots, N\},$$

where $B$ is a constant with the dimension of frequency. For this model they obtain the autocorrelation function

$$E[\varepsilon_t \varepsilon_0^*] = \sum_{n=1}^{N} E[a_n^2] e^{-B|t|/2} \int_0^{2\pi} e^{i2\pi\, \mathrm{f_D}\cos(\alpha) t} d\alpha. \tag{25}$$

For our model in the setting (22), we recall from (23) that the autocorrelation is given by the quite similar expression

$$E[Z_t Z_0^*] = \frac{e^{-C|t|}}{2} \int_0^{2\pi} (a^+(\alpha))^2 e^{i2\pi\, \mathrm{f_D}\, \cos(\alpha) t}\, p(\alpha)\, d\alpha.$$

**Remark 4.1.** *It should be noted that the autocorrelation similarity is obtained although the modeling assumptions are quite different: in our MFC model the amplitude is governed by a Poisson flip process, Feng and Field's model has time invariant amplitude functions; our model updates a phase when the corresponding amplitude flips, their model has Wiener process phase evolution.*

4.2. **Model parameter estimation from PSD measurements.** Feng and Field's publication [9] presents a method for estimating the flip rate constant $C$ from measurements which can be used when the scaled scatterer density $(a^+)^2 p$ is known. Namely, given a measurement of the PSD, $\widehat{S}(f)$, estimate $C$ by

$$C = \arg\min_{x>0} \cos^{-1}\left( \int_{\mathbb{R}} \sqrt{\widehat{S}(f) S_x(f)}\, df \right), \tag{26}$$

where the PSDs are scaled so that $\|\widehat{S}\|_1 = 1$ and $\|S_x\|_1 = 1$. However, from (24) we see that the shape of the PSD depends both on the flip rate and the scaled scatterer density on $(a^+)^2 p$, see Figure 5. So to estimate the flip rate from PSD measurements, $(a^+)^2 p$ either has to be known or it has to be estimated. It is more difficult and costly to estimate $(a^+)^2 p$ than the flip rate, we restrict ourselves to a tentative approach to this problem. Let $\widehat{S}$ denote a real life PSD measurement for a moving receiver, and consider relation (24) as the inhomogeneous Fredholm integral equation

$$\int_0^{2\pi} \mathcal{K}_C(f,\alpha)(a^+(\alpha))^2 p(\alpha) d\alpha = \widehat{S}(f),$$

with unknown $(a^+(\alpha))^2 p(\alpha)$ and parametrized by the flip rate $C$. For a fixed flip rate, this problem might be discretized to a system of linear equations and, if expecting noise in the measurement $\widehat{S}$, solved as an inverse problem using Tikhonov or similar regularization techniques, c.f. [14]. For example, if we denote the discretized integral kernel $\mathcal{K}_C \in \mathbb{R}^{m\times n}$, then the Tikhonov regularized solution has the variational representation

$$(a_C^+)^2 p_C = \arg\min_{x\in\mathbb{R}^n} \|\mathcal{K}_C x - \widehat{S}\|^2 + \beta \|x\|^2 \tag{27}$$

for a regularizing parameter $\beta > 0$ to be chosen. If the flip rate $C$ and $(a^+(\alpha))^2 p(\alpha)$ have to be determined simultaneously, we suggest coupling the solution of (27) with a minimization problem on the form

$$C = \arg\min_{x>0} \|\mathcal{K}_x (a_x^+)^2 p_x - \widehat{S}\|.$$

4.3. **A PSD based Gaussian process algorithm.** We now present a PSD based algorithm for generating signal realizations of WSS complex Gaussian processes for the modeling the setting (22). The PSD (24) can be used to represent the Gaussian process spectrally by the Itô integral

$$Z_t = \int_{\mathbb{R}} e^{i2\pi f t} \sqrt{S_C(f)}\, dW_f, \tag{28}$$

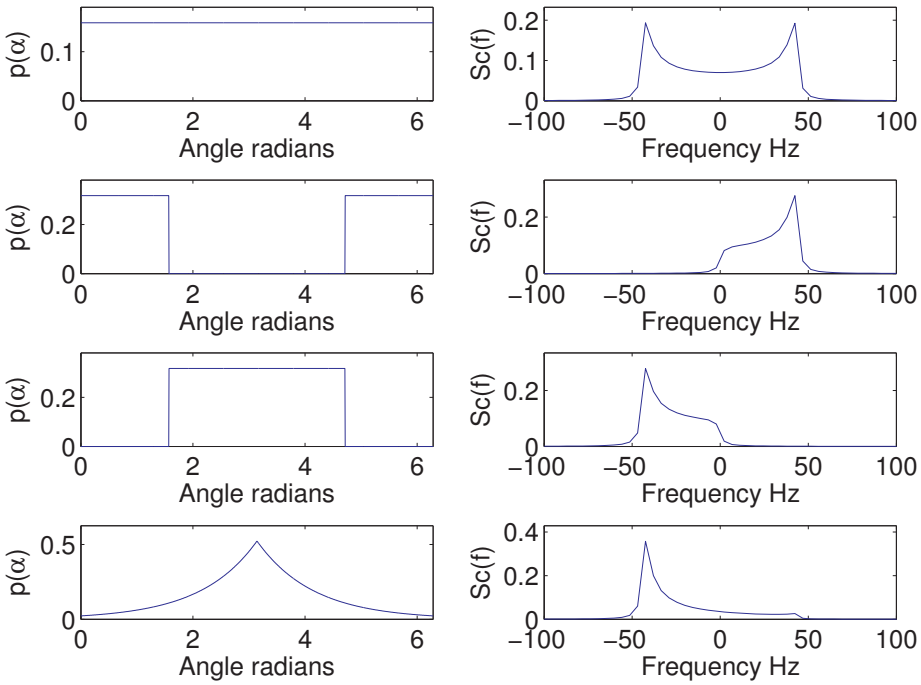


FIGURE 5. Right column plots illustrate the PSD $S_C(f)$ obtained by equation (24) when the scaled scatterer density $(a^+)^2 p(\alpha)$ is given by respective left column plots, $C = 10$, and $\mathrm{f_D} = 50$. **Top row:** $(a^+)^2 p = 1/(2\pi)$ yields the mollified Jakes' spectrum PSD. **Second row:** $(a^+)^2 p(\alpha) = 1_{[-\pi/2,\pi/2]}(\alpha)/\pi$ yields a PSD consisting almost exclusively of positive Doppler shifts since the receiver's moves towards the active scatterers. **Third row:** $(a^+)^2 p(\alpha) = 1_{[\pi/2,3\pi/2]}(\alpha)/\pi$ yields a PSD consisting almost exclusively of negative Doppler shifts since the receiver moves away from the active receivers. **Last row:**$(a^+)^2 p(\alpha) = \exp(-|\alpha - \pi|)/scale$.

where $W_f \in \mathbb{C}$ is a complex Wiener processes with independent infinitesimal increments $dW_f \sim \mathcal{N}_{\mathbb{C}}(0, df)$. Seeking to approximate the integral (28) by quadrature we first approximate the integrand by a compactly supported function by choosing a cut off frequency $\mathrm{f_M} \gg \mathrm{f_D}$ and introducing the $C^3(\mathbb{R})$ window function

$$\varrho(f) = \begin{cases} 0 & \text{if } f \leq -\mathrm{f_M} \\ \varrho_I(2(\mathrm{f_M} + f)/\mathrm{f_M}) & \text{if } -\mathrm{f_M} < f < -\mathrm{f_M}/2 \\ 1 & \text{if } |f| \leq \mathrm{f_M}/2 \\ \varrho_I(2(\mathrm{f_M} - f)/\mathrm{f_M}) & \text{if } \mathrm{f_M}/2 < f < \mathrm{f_M} \\ 0 & \text{else,} \end{cases} \tag{29}$$

where $\varrho_I(f) = -20f^7 + 70f^6 - 84f^5 + 35f^4$ is the unique solution of the Birkhoff-Hermite interpolation with conditions $\varrho_I^{(j)}(0) = 0$ for $j = 0, 1, \ldots, 4$ and $\varrho_I(1) = 1$,

$\varrho_I^{(j)}(1) = 0$ for $j = 1, \ldots, 4$. Replacing the the integrand term $\sqrt{S_C(f)}$ in (28) with $\sqrt{\varrho(f)S_C(f)}$ yields the integral

$$\int_{-\mathrm{f_M}}^{\mathrm{f_M}} e^{i2\pi ft}\sqrt{\varrho(f)S_C(f)}dW_f. \tag{30}$$

Quadrature approximating this integral with the Inverse Discrete Fourier Transform (IDFT) gives output signal realizations by

$$\overline{\mathrm{Z}}_t = \sum_{j=1}^{M_1} e^{i2\pi f_j t}\sqrt{\varrho(f_j)\overline{S}_C(f_j)}\Delta W_j. \tag{31}$$

Here $f_j = -\mathrm{f_M} + (j-1)\Delta f$, $\Delta f = 2\,\mathrm{f_M}/(M_1 - 1)$, $\Delta W_j \sim \mathcal{N}_{\mathbb{C}}(0, \Delta f)$ are i.i.d. complex Wiener increments and $\overline{S}_C(f_j)$ approximates $S_C(f_j)$ by quadrature of the integral (24),

$$\overline{S}_C(f_k) = \sum_{j=1}^{M_2} \frac{C(a^+(x_j))^2 p(x_j)\tilde{\nu}_j}{C^2 + (2\pi(f_k - \mathrm{f_D}\cos(x_j)))^2} \tag{32}$$

on a grid $0 = x_1 < x_2 < \ldots < x_{M_2} = 2\pi$ with integration weights $\tilde{\nu}_\cdot \geq 0$ which satisfies $\sum_{j=1}^{M_2} \tilde{\nu}_j = 2\pi$.

Algorithms for generating Gaussian process signal realizations using the IDFT have been described in [12, 15], and in Algorithm 4, we present a version suited for our setting.

**Algorithm 4** A PSD based Gaussian process algorithm

**Input:** Flip rate $C$, maximum Doppler shift $\mathrm{f_D}$, spectral cutoff $\mathrm{f_M}$, sampling times $\mathbf{t} = (t_1, t_2, \ldots, t_N)$ and scatterer density $p$.
**Output:** Gaussian process realization $\{\overline{\mathrm{Z}}_{t_j}\}_{j=1}^N$.
Construct a grid $f_j = -\mathrm{f_M} + (j-1)\Delta f$, for $j = 1, 2, \ldots, M_1$ with $\Delta f = 2\,\mathrm{f_M}/(M_1 - 1)$.
Construct a grid $0 = \alpha_1 < \alpha_2 < \ldots < \alpha_{M_2} = 2\pi$ and quadrature weights $\tilde{\nu}_\cdot \geq 0$ for which $\sum_{j=1}^{M_2} \tilde{\nu}_j = 2\pi$.
**for** $j = 1$ to $M_1$ **do**
  Compute $\overline{S}_C(f_j)$ according to (32).
**end for**
Generate i.i.d. Wiener increments $\{\Delta W_j\}_{j=1}^{M_2}$ distributed according to $\Delta W_j \sim \mathcal{N}_{\mathbb{C}}(0, \Delta f)$.
**for** $k = 1$ to $N$ **do**
  Compute $\overline{\mathrm{Z}}_{t_k}$ by the IDFT (31).
**end for**

4.4. **Computational cost of Algorithm 4.** The error analysis for this algorithm is quite similar to the error analysis of Algorithm 2 in the sense that the distributional error can be bounded in terms of the difference between the covariance matrix of realizations of Algorithm 4, denoted $\overline{\mathrm{K}}$, and the covariance matrix $K$ for the sampled limit complex Gaussian process $Z_{\mathbf{t}}$ derived in (12). Our procedure for obtaining an error bound is to first derive a bound of $\|\,\overline{\mathrm{K}} - K\|_2$ and thereafter use

Theorem 3.10 to bound $e(\overline{\mathrm{Z}}_{\mathbf{t}})$ from above. We start with computing the elements of $\overline{\mathrm{K}}$.

The representation (31) implies that for two arbitrary times $t_j, t_k$ from the sampling times $\mathbf{t} = (t_1, t_2, \ldots, t_N)$,

$$\begin{aligned}
\overline{\mathrm{K}}(t_j, t_k) &= E\left[\overline{\mathrm{Z}}_{t_j}\, \overline{\mathrm{Z}}^*_{t_k}\right] \\
&= \sum_{l,m=1}^{M_1} e^{i2\pi(f_l t_j - f_m t_k)} \sqrt{\overline{S}_C(f_l)\overline{S}_C(f_m)} \sqrt{\varrho(f_l)\varrho(f_m)} E\left[\Delta W_l \Delta W_k^*\right] \\
&= \sum_{l=1}^{M_1} e^{i2\pi f_l(t_j - t_k)} \varrho(f_l)\overline{S}_C(f_l)\Delta f,
\end{aligned}$$

where we recall that $\overline{S}_C$ is an approximation of the PSD $S_C$. The error of the covariance terms is split into three parts,

(33)

$$\begin{aligned}
|K(t_j, t_k) - \overline{\mathrm{K}}(t_j, t_k)| &= \left| \int_{\mathbb{R}} e^{i2\pi f(t_j - t_k)} S_C(f)\, df - \sum_{l=1}^{M_1} e^{i2\pi f_l(t_j - t_k)} \varrho(f_l)\overline{S}_C(f_l)\Delta f \right| \\
&\leq \left| \int_{\mathbb{R}} e^{i2\pi f(t_j - t_k)} (1 - \varrho(f)) S_C(f)\, df \right| \\
&+ \left| \int_{-\mathrm{f_M}}^{\mathrm{f_M}} e^{i2\pi f(t_j - t_k)} \varrho(f) S_C(f)\, df - \sum_{l=1}^{M_1} e^{i2\pi f_l(t_j - t_k)} \varrho(f_l) S_C(f_l)\nu_l \right| \\
&+ \left| \sum_{l=1}^{M_1} e^{i2\pi f_l(t_j - t_k)} \varrho(f_l)(S_C(f_l) - \overline{S}_C(f_l))\nu_l \right| \\
&= I + II + III.
\end{aligned}$$

Under the assumption $C = O(1)$, we see from equation (24) that $S_C(f) = O((1 + f^2)^{-1})$. Therefore $I = O(\mathrm{f_M}^{-1})$. For the second term [4, thm. 6.3] implies that since $\varrho S_C$ up third derivative is continuous and (can be considered) $2\,\mathrm{f_M}$ periodic on $[-\mathrm{f_M}, \mathrm{f_M}]$, $II = O((\mathrm{f_M}/M_1)^4)$, where the implicit constant in the error term depends on the derivatives of $\varrho S_C$ up to third order. (Increasing the number of derivative conditions in the Birkhoff interpolation $\varrho_I$, will increase the convergence order in $II$, but, as a trade off, the error term's implicit constant increases as well.) For the last term, the approximation $\overline{S}_C(f)$ is obtained from the quadrature (32) using $M_2$ integration points, so that $|S_C(f) - \overline{S}_C(f)| = O((f^2+1)^{-1} M_2^{-\gamma})$, where the convergence order $\gamma > 0$ depends on the numerical integrator used and is the same order as that obtained in (15) for Algorithm 2. Thereby

$$III \leq \sum_{l=1}^{M_1} |S_C(f_l) - \overline{S}_C(f_l)|\Delta f = O\left( \int_{-\mathrm{f_M}}^{\mathrm{f_M}} (f^2+1)^{-1} df M_2^{-\gamma} \right) = O(M_2^{-\gamma}). \tag{34}$$

Adding the three error terms yields the covariance error bound

$$|K(t_j, t_k) - \overline{\mathrm{K}}(t_j, t_k)| \leq \varepsilon = O\left(\mathrm{f_M}^{-1} + (\mathrm{f_M}/M_1)^4 + M_2^{-\gamma}\right), \tag{35}$$

and by applying Theorem 3.10 we obtain the following error bound for this algorithm.

**Corollary 4.2.** *Let* $\overline{\mathrm{K}}$ *denote the covariance matrix of the stochastic process generated by Algorithm 4 in the modeling setting* (22) *and assume the covariance matrix* $K$ *of the Gaussian process* $Z_{\mathbf{t}}$*, as given by* (12)*, is non-singular and representable by the SVD* $K = USU^H$ *with* $S = diag(s_i)$*,* $s_1 \geq s_2 \geq \ldots \geq s_N > 0$*. Then, if* $M_1$*,* $M_2$ *and* $\mathrm{f_M}$ *are chosen sufficiently large so that the error bound* (35) $\|K - \overline{\mathrm{K}}\|_2 \leq \sqrt{N}\epsilon$ *is fulfilled with* $10N^{3/2}\epsilon < s_N$*, realizations of Algorithm 4*

$$e(\overline{\mathrm{Z}}_{\mathbf{t}}) = O\left(\frac{N^{3/2}\left({\mathrm{f_M}}^{-1} + (\mathrm{f_M}/M_1)^4 + M_2^{-\gamma}\right)}{s_N}\right).$$

*Proof.* The result follows from Theorem 3.10. □

To fulfill the accuracy condition $e(\overline{\mathrm{Z}}_{\mathbf{t}}) \leq \mathrm{TOL}$, Corollary 4.2 implies that $\mathrm{f_M} = O\left((S_N\ \mathrm{TOL})^{-1}N^{3/2}\right)$, $M_1 = O\left(((S_N\ \mathrm{TOL})^{-1}N^{3/2})^{5/4}\right)$ and $M_2 = O\left(((S_N\ \mathrm{TOL})^{-1}N^{3/2})^{1/\gamma}\right)$. The cost of generating one signal realization for this algorithm is the sum of the cost of computing $\overline{S}_C(f_k)$ for $k = 1, 2, \ldots, M_1$ by (32), which generally ammounts to $O(M_1M_2)$, and the cost $O(NM_1)$ for computing (31) to generate a realization $\overline{\mathrm{Z}}_{\mathbf{t}}$. The cost of generating $L$ signal realizations thus becomes

$$\mathrm{Cost(Algorithm\ 4)} = O\left(\left(\frac{N^{3/2}}{S_N\ \mathrm{TOL}}\right)^{5/4+1/\gamma} + LN\left(\frac{N^{3/2}}{S_N\ \mathrm{TOL}}\right)^{5/4}\right).$$

**Remark 4.3.** *In many settings, particularly the setting when generating realizations on uniformly sampled grid points on an interval* $[0, T)$ *with* $\Delta t = T/N$*, it is possible to apply the FFT techniques to speed up the quadrature computations of the discrete convolution* (32) *and the IDFT* (31) *so that the cost of generating* $L$ *realizations with Algorithm 4 in terms of* $L, N$ *and* TOL *instead amounts to*

$$\mathrm{Cost(Alg\ 4, FFT)} = O\left(\left(L\left(\frac{N^{3/2}}{S_N\,\mathrm{TOL}}\right)^{5/4} + \left(\frac{N^{3/2}}{S_N\,\mathrm{TOL}}\right)^{1+1/\min(4,\overline{\gamma})}\right)\log\left(\frac{N^{3/2}}{S_N\,\mathrm{TOL}}\right)\right).$$

*Here* $\overline{\gamma}$ *is the convergence order of the discrete convloution* (32) *which has the relation* $\overline{\gamma} \leq \gamma$ *since when computing* (32) *with FFT, the quadrature grid has to fulfill* $\cos(x_{j+1}) - \cos(x_j) = \Delta f/\,\mathrm{f_D}$*, which generally does not optimize the accuracy of the computation.*

## 5. Summary of the complexity estimates

Having estimated upper bounds for the computational cost of generating output realizations by four different algorithms, we now summarize the results. We do however stress that none of the error bounds for which the cost estimates are based are proven to be sharp, so the following cost comparison should not be considered conclusive. In Table 1 we present cost estimates for non-WSS modeling settings, which occurs if the amplitude function $a^+$ is time dependent or/and if modeling with the full delay function $\tau(\alpha, t)$. The results of the table indicates that for settings when $\gamma \geq 3/5$, the covariance based Gaussian process algorithm, Algorithm 2, is asymptotically the most efficient algorithm.

In modeling settings resulting in WSS output processes,

$$\tau(\alpha, t) = -v\cos(\alpha)t/c, \quad \text{and} \quad a^+(\alpha, t) = a^+(\alpha), \tag{36}$$

we derived the computational cost presented in Table 2. The results of the table indicate that in the WSS setting, Algorithm 3 outperforms the other algorithms in terms of efficiency.

| | Computational cost |
|---|---|
| Algorithm 1 | $O\left(L\,\mathrm{TOL}^{-2}\left(\frac{N^{3/2}}{S_N}\right)^3\right)$ |
| Algorithm 2 | $O\left(N^3 + N^2\left(\frac{N^{3/2}}{s_N\,\mathrm{TOL}}\right)^{1/\gamma} + LN^2\right)$ |

TABLE 1. Computational cost for non-WSS modeling settings.Recall that $\gamma$ is the convergence order for the quadrature (15).

| | Computational cost |
|---|---|
| Alg 1 | $O\left(L\,\mathrm{TOL}^{-2}\left(\frac{N^{3/2}}{S_N}\right)^3\right)$ |
| Alg 2 | $O\left(N^3 + N\left(\frac{N^{3/2}}{s_N\,\mathrm{TOL}}\right)^{1/\gamma} + LN^2\right)$ |
| Alg 3 | $O\left(N\left[\frac{N^{1/2}}{\mathrm{TOL}}\min\left(\frac{N}{s_N}, \mathrm{tr}(\Lambda^{-1})\right)\right]^{1/\gamma} + LN\log(N)\right)$ |
| Alg 4 | $O\left(\left(\frac{N^{3/2}}{S_N\,\mathrm{TOL}}\right)^{5/4+1/\gamma} + LN\left(\frac{N^{3/2}}{S_N\,\mathrm{TOL}}\right)^{5/4}\right)$ |
| Alg 4, FFT | $O\left(\left(L\left(\frac{N^{3/2}}{S_N\,\mathrm{TOL}}\right)^{5/4} + \left(\frac{N^{3/2}}{S_N\,\mathrm{TOL}}\right)^{1+1/\min(4,\overline{\gamma})}\right)\log\left(\frac{N^{3/2}}{S_N\,\mathrm{TOL}}\right)\right)$ |

TABLE 2. Computational cost for the WSS setting (36). Recall that $\gamma$ is the quadrature convergence order obtained in (15) and (34) for the respective algorithms, and $\overline{\gamma}$ with $\overline{\gamma} \leq \gamma$ is the convergence order obtained when computing (34) with FFT.

For better understanding of the cost estimates, we would like an estimate on how $N/S_N$ and $\mathrm{tr}(\Lambda^{-1})$ depend on the sample times $\mathbf{t} = (0, t_2, \ldots, t_N)$ used in the generation of signal realizations. A general answer to this question is however too demanding; we restrict ourselves to the WSS modeling setting (22).

The eigenvalue $\Lambda_n$ is a DFT approximation of $S_C((n-1)/T)$, and we may, by introducing a cutoff function as in (29) and arguing similar as in the error analysis of (33), derive the error bound

$$|\Lambda_n - S_C((n-1)/t_N)| \leq O(\Delta t^4 + e^{-C|t_N|}), \quad n = 1, 2, \ldots, N,$$

and

$$|\Lambda_{N+n} - S_C(-(N-1-n)/t_N)| \leq O(\Delta t^4 + e^{-C|t_N|}), n = 1, 2, \ldots, N-2.$$

Then it follows that for settings $(4/C)\log(N) \leq t_N = o(N)$ and $\Delta t = t_N/(N-1)$, we obtain the sharp bound

$$\begin{aligned}\sum_{j=1}^{2N-2} \Lambda_j^{-1} &= O\left(\frac{t_N^2}{N}\right) + \sum_{j=2-N}^{N-1} S_C(j/t_N)^{-1} \\ &= O\left(\frac{t_N^2}{N} + \sum_{j=2-N}^{N-1} (j/t_N)^2\right) \\ &= O\left(\frac{N^3}{t_N^2}\right),\end{aligned} \tag{37}$$

Figure 6 illustrates this bound for a numerical example. To relate the magnitude of

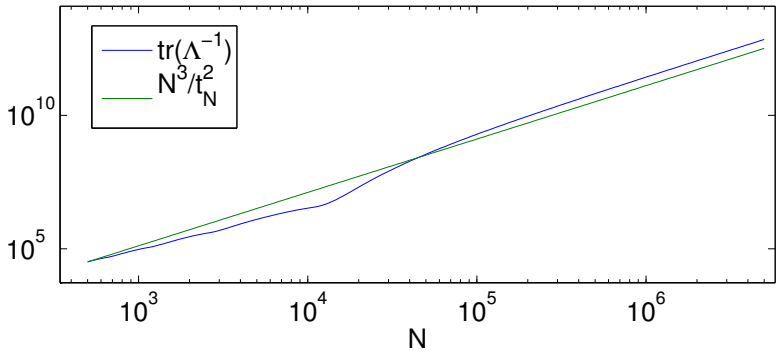


FIGURE 6. Numerical study of how $\mathrm{tr}(\Lambda^{-1})$ varies as a function of $N$ in the modeling setting $C = 10$, $\mathrm{f_D} = 50\mathrm{Hz}$, $(a^+)^2 p = \pi^{-1}$, $t_N = \sqrt{N}$ and $N$ uniform sample times on the interval $[0, t_N)$.

$s_N$ to $\Lambda$, we note that since $K$ is circulantly embedded into $\mathcal{K}$, Cauchy's interlacing theorem says that $\min_j \Lambda_j \leq s_N \leq \tilde{\Lambda}_N$ with $\tilde{\Lambda}_j$ denoting the $j$th largest eigenvalue of $\Lambda$. In settings where estimate (37) is valid, we thereby derive the upper bound $N/s_N = O(N^3/t_N^2)$, giving the relation $N/s_N = O(\mathrm{tr}(\Lambda^{-1}))$. We end these informal estimates with the numerical study in Figure 7, showing a setting where $N/s_N$ is orders of magnitude smaller than $\mathrm{tr}(\Lambda^{-1})$.

## 6. Numerical examples

6.1. **Example 1.** The first numerical example compares realizations generated by Algorithm 1 and 2 in a WSS model setting with the scaled scatterer density $(a^+)^2 p(\alpha) = 1/\pi$ and the delay function $\tau(\alpha, t) = -v\cos(\alpha)t/c$. The flip rate is determined from a real life signal measurement from a receiver moving with the speed $v = 6.9\mathrm{m/s}$ and carrier frequency $\mathrm{f_c} = 1.8775\mathrm{GHz}$ sampled uniformly the time interval $[-0.08, 0.08]$ using $N = 1419$ samples, and signal realizations are thereafter generated using the same modeling parameters. From the measured signal $\widehat{Z}(t)$, we approximate the PSD by the FFT of $\widehat{Z}(t)\widehat{Z}(0)^* =: \widehat{S}(f)$ and determine the flip

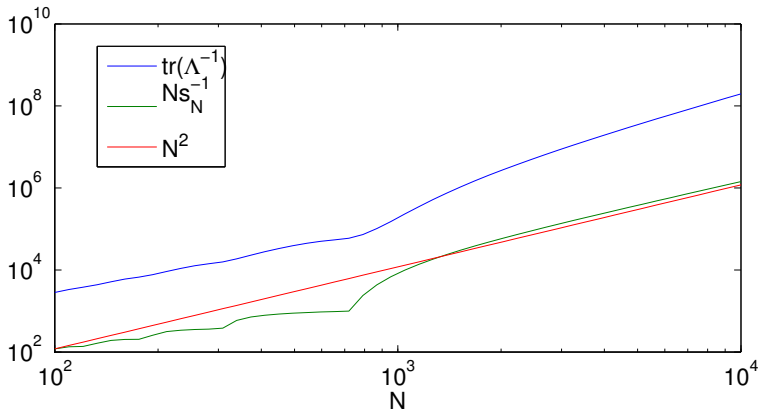


FIGURE 7. Numerical comparison of $N/s_N$ and $\mathrm{tr}(\Lambda^{-1})$ in the modeling setting $C = 15$, $f_D = 50$Hz, $(a^+)^2 p = \pi^{-1}$, $t_N = \log(N)$ and $N$ sample uniform sample times on the $[0, t_N)$. (Since $s_N$ is computed by the SVD, it cannot be computed for as large $N$ values as $\mathrm{tr}(\Lambda)$, cf. Figure 6.)

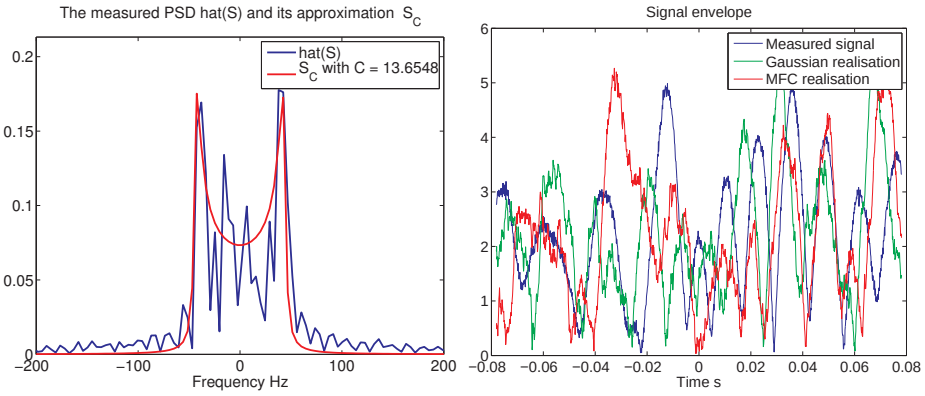


FIGURE 8. **Left plot:** The best fit of $S_C(f)$ to the measured signal's PSD $\widehat{S}(f)$ as a function of the flip rate $C$, which is determined by (38). **Right plot:** Scaled signal envelopes of a measured signal, a signal realization from the MFC algorithm, and a realization from the Gaussian process algorithm.

rate by

$$C = \arg\min_{\widetilde{C}>0} \|(\widehat{S} - S_{\widetilde{C}})\varrho\|_1, \tag{38}$$

which is a slight modification of Feng and Field's idea (26). See Figure 8 for an comparison of $S_C$ and $\widehat{S}$ for the best fit flip rate $C \approx 13.65$.

Having determined the flip rate, $L = 2000$ signal realizations are generated by the MFC algorithm using $M = 2000$ scatterers and for the covariance based

Gaussian Process algorithm, one may derive using pen and paper that $K(t_j, t_k) = J_0(2\pi\,\mathrm{f_D}(t_k - t_j))$, $J_0$ denoting the Bessel function of the first kind. Autocorrelation and PSD functions computed for respective algorithms are plotted in figures 8 and 9. For signal realizations of Algorithm 1, the autocorrelation is approximated by taking the sample average of

$$E[\overline{\mathrm{Z}}_t\,\overline{\mathrm{Z}}_0^*] \approx \sum_{j=1}^{L} \frac{\overline{\mathrm{Z}}_t(\omega_j)\,\overline{\mathrm{Z}}_0(\omega_j)^*}{L} \tag{39}$$

with $\overline{\mathrm{Z}}_t(\omega_j)$ denoting the $j$th signal realization. The approximation 39 is compared with the autocorrelation of signal realizations of Algorithm 2 (which is given by the first row of the covariance matrix $K$).

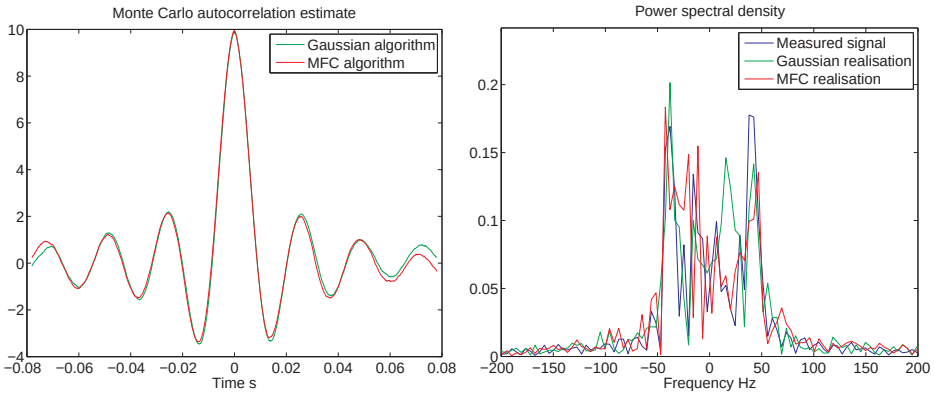


FIGURE 9. **Left plot:** Monte Carlo estimated autocorrelation for realizations generated by Algorithm 1, cf. (39), and the autocorrelation for Gaussian process signal realizations $A(t) = J_0(2\pi\,\mathrm{f_D}\,t)$. **Right plot:** Scaled PSD of the measured signal and of one signal realization from each the algorithms studied.

6.2. **Example 2.** In the second example, we model the temporally varying scattering environment with a mobile receiver moving from left to right through a thin opening of a non-reflecting wall, as sketched in Figure 10. When situated on the left side of the opening, the mobile receiver receives scattered rays at its rear, and when the mobile receiver is on the right side of the opening, it receives rays at its front. As a model for this change in scattering environment we consider the the time interval $[0, 2)$ seconds, assume that the receiver moves through the opening at $t = 1$ and set

$$a^+(\alpha, t) = \begin{cases} \cos^2(\alpha) 1_{(\pi/2, 3\pi/2)}(\alpha) & \text{for } 0 \le t < 1 \\ \cos^2(\alpha) 1_{(-\pi/2, \pi/2)}(\alpha) & \text{for } 1 \le t \le 2. \end{cases} \tag{40}$$

Other modeling parameters are set to $p = (2\pi)^{-1}$, $C = 15$ and $f_D = 43.5$Hz. Figure 11 contains snapshots of the time dependent PSD for a single stochastic signal realization created by Algorithm 2. It shows that when the mobile receiver is situated on the left side of the opening, the the PSD is concentrated around $-f_D$, and when the receiver is on the right side of the opening, the PSD is concentrated

around $f_D$. By the discussion of the relation between $(a^+)^2 p$ and $S_C(f)$ given in Section 4, the snapshotted PSDs in Figure 11 are reasonable.

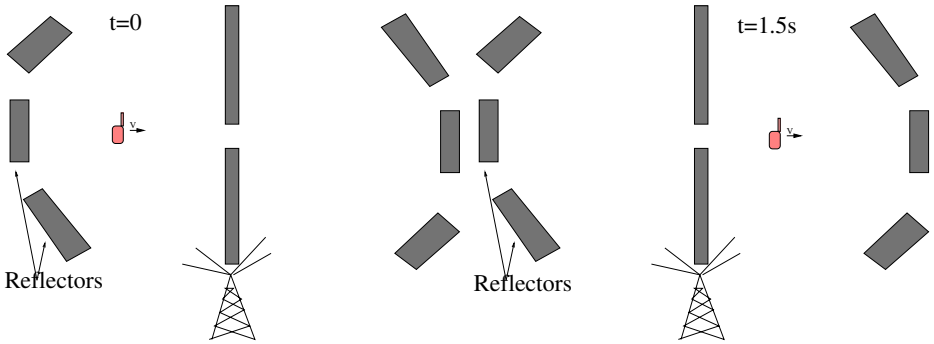


FIGURE 10. A receiver moving rightwards through a thin opening in a non-scattering wall and thereby experiencing a change in the scattering environment.

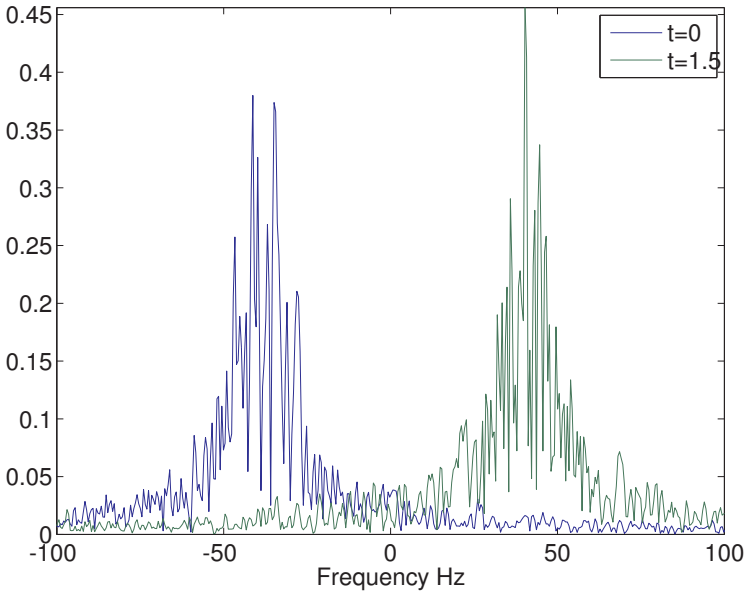


FIGURE 11. Snapshots of the time dependent PSD for a signal realization at $t = 0.5$s when the receiver is situated on the left side of the wall opening (blue line) and at $t = 1.5$s situated at to the right side of the wall opening (green line).

## Appendix A. Theorems

**Corollary A.1** ([8][p. 198]). *Let $G$ and $F$ be arbitrary matrices (of the same size) where $\sigma_1 \geq \ldots \geq \sigma_n$ are the singular values of $G$ and $\sigma'_1 \geq \ldots \geq \sigma'_n$ are the singular values of $G + F$. Then $|\sigma_i - \sigma'_i| \leq \|F\|_2$.*

## Appendix B. Approximation of the delay function

The delay function primarily considered in this paper, $\tau(\alpha, t) = -vt\cos(\alpha)/c$, is a first order approximation of the full delay function. Here we describe how this approximation is obtained.

Assuming the scattering boundary is described by $\{(\alpha, R(\alpha))|0 \leq \alpha \leq 2\pi\}$ and that the receiver is moving in the direction $(v, 0)$, the analytical delay function is given by

$$\begin{aligned}\tau(\alpha, t) &= \frac{\sqrt{(R(\alpha)\cos(\alpha) - vt)^2 + R(\alpha)^2\sin(\alpha)^2}}{c} \\ &= \frac{\sqrt{(vt)^2 - 2vtR(\alpha)\cos(\alpha) + R(\alpha)^2}}{c}.\end{aligned}$$

A Taylor expansion of this function with respect to $vt$ yields

$$\tau(\alpha, t) = \frac{R(\alpha)}{c} - \frac{\cos(\alpha)}{c}vt + O\left(\frac{vt}{cR(\alpha)}\right).$$

Assuming that for the times of interest $vt/(R(\alpha)c)$ is a small term, a good approximation of the delay function as a function of time is

$$\tau(\alpha, t) = -\frac{\cos(\alpha)}{c}vt.$$

Here the term $R(\alpha)/c$ is removed since it is constant with respect to time and thus can be considered as part of the random phase shift term $\theta_\alpha$ of $\exp(-i(2\pi\,\mathrm{f_c}\,\tau(\alpha, t) + \theta_\alpha))$. As this first order approximation considers the Doppler effect from a given scatterer to be constant, it is only valid when the receiver is far away from the scattering boundary.

**Acknowledgement.** The authors would like to thank professor Anders Szepessy for fruitful discussions and Henrik Asplund at Ericsson Research for providing signal measurements and modeling extension ideas.

## References

[1] TR 25.996 3GPP. *3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Spatial channel model for Multiple Input Multiple Output (MIMO) simulations (Release 6)*. 3GPP, 2003.

[2] Søren Asmussen and Peter W. Glynn. *Stochastic simulation: algorithms and analysis*, volume 57 of *Stochastic Modelling and Applied Probability*. Springer, New York, 2007.

[3] V. Bentkus. On the dependence of the Berry-Esseen bound on dimension. *J. Statist. Plann. Inference*, 113(2):385–402, 2003.

[4] William L. Briggs and Van Emden Henson. *The DFT*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1995. An owner's manual for the discrete Fourier transform.

[5] Charalambos D. Charalambous, Seddik M. Djouadi, and Christos Kourtellaris. Statistical analysis of multipath fading channels using generalizations of shot noise. *EURASIP J. Wirel. Commun. Netw.*, 2008:1–9, 2008.

[6] Chia-Chin Chong, Chor-Min Tan, D.I. Laurenson, S. McLaughlin, M.A. Beach, and A.R. Nix. A novel wideband dynamic directional indoor channel model based on a markov process. *Wireless Communications, IEEE Transactions on*, 4(4):1539 – 1552, jul. 2005.
[7] R. H. Clarke. A statistical theory of mobile-radio reception. *Bell Sys. Tech.*, 47:957–1000, 1968.
[8] James W. Demmel. *Applied numerical linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.
[9] Tao Feng and T.R. Field. Statistical analysis of mobile radio reception: an extension of clarke's model. *Communications, IEEE Transactions on*, 56(12):2007–2012, december 2008.
[10] Tao Feng, T.R. Field, and S. Haykin. Stochastic differential equation theory applied to wireless channels. *Communications, IEEE Transactions on*, 55(8):1478 –1483, aug. 2007.
[11] Markos A. Katsoulakis and Anders Szepessy. Stochastic hydrodynamical limits of particle systems. *Commun. Math. Sci.*, 4(3):513–549, 2006.
[12] J.I. Smith. A computer generated multipath fading simulation for mobile radio. *Vehicular Technology, IEEE Transactions on*, 24(3):39 – 40, aug 1975.
[13] N. G. van Kampen. *Stochastic processes in physics and chemistry*, volume 888 of *Lecture Notes in Mathematics*. North-Holland Publishing Co., Amsterdam, 1981.
[14] Curtis R. Vogel. *Computational methods for inverse problems*, volume 23 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. With a foreword by H. T. Banks.
[15] D.J. Young and N.C. Beaulieu. The generation of correlated rayleigh random variates by inverse discrete fourier transform. *Communications, IEEE Transactions on*, 48(7):1114 – 1127, jul 2000.
[16] T. Zwick, C. Fischer, and W. Wiesbeck. A stochastic multipath channel model including path directions for indoor environments. *Selected Areas in Communications, IEEE Journal on*, 20(6):1178 – 1192, aug. 2002.

Håkon Hoel, Department of Numerical Analysis and Computer Science, KTH, SE-100 44, Stockholm, Sweden.
*E-mail address*: haakonah1@gmail.com

Henrik Nyberg, Ericsson Research, SE-164 80, Stockholm, Sweden.
*E-mail address*: Henrik.L.Nyberg@ericsson.com

# Paper II

# Implementation and Analysis of an Adaptive Multilevel Monte Carlo Algorithm

Håkon Hoel[*], Erik von Schwerin[†], Anders Szepessy[‡], and Raúl Tempone[§]

April 23, 2012

**Abstract**

This work generalizes a multilevel Monte Carlo (MLMC) method introduced in [7] for the approximation of expected values of functions depending on the solution to an Itô stochastic differential equation. The work [7] proposed and analyzed a forward Euler MLMC method based on a hierarchy of uniform time discretizations and control variates to reduce the computational effort required by a standard, single level, forward Euler Monte Carlo method from $\mathcal{O}\left(\epsilon^{-3}\right)$ to $\mathcal{O}\left((\epsilon^{-1}\log(\epsilon^{-1}))^2\right)$ for a mean square error of size $\epsilon^2$. This work uses instead a hierarchy of adaptively refined, non uniform, time discretizations, generated by an adaptive algorithm introduced in [20, 19, 5]. Given a prescribed accuracy TOL in the weak error, this adaptive algorithm generates time discretizations based on a posteriori expansions of the weak error first developed in [24]. A theoretical analysis gives results on the stopping, the accuracy, and the complexity of the resulting adaptive MLMC algorithm. In particular, it is shown that: the adaptive refinements stop after a finite number of steps; the probability of the error being smaller than TOL is under certain assumptions controlled by a given confidence parameter, asymptotically as $\mathrm{TOL} \to 0$; the complexity is essentially the expected for MLMC methods, but with better control of the constant factors. We also show that the multilevel estimator is asymptotically normal using the Lindeberg-Feller Central Limit Theorem. These theoretical results are based on previously developed single level estimates, and results on Monte Carlo stopping from [3]. Our numerical tests include cases, one with singular drift and one with stopped diffusion, where the complexity of uniform single level method is $\mathcal{O}\left(\mathrm{TOL}^{-4}\right)$. In both these cases the results confirm the theory by exhibiting savings in the computational cost to achieve an accuracy of $\mathcal{O}\left(\mathrm{TOL}\right)$, from $\mathcal{O}\left(\mathrm{TOL}^{-3}\right)$ for the adaptive single level algorithm to essentially $\mathcal{O}\left(\mathrm{TOL}^{-2}\log\left(\mathrm{TOL}^{-1}\right)^2\right)$ for the adaptive MLMC.

[*]CSC, Royal Institute of Technology (KTH), Stockholm, Sweden

[†]Applied Mathematics and Computational Sciences, KAUST, Thuwal, Saudi Arabia

[‡]Mathematics, Royal Institute of Technology (KTH), Stockholm, Sweden

[§]Applied Mathematics and Computational Sciences, KAUST, Thuwal, Saudi Arabia (raul.tempone@kaust.edu.sa)

**Key words**: computational finance, Monte Carlo, multi-level, adaptivity, weak approximation, error control, Euler–Maruyama method, a posteriori error estimates, backward dual functions, adjoints

**AMS subject classification**: 65C30, 65Y20, 65L50, 65H35, 60H35, 60H10

# Contents

**1 Introduction** **2**
1.1 A single level posteriori error expansion . . . . . 6

**2 Adaptive Algorithms and Multilevel Variance Reduction** **10**
2.1 Path independent time stepping . . . . . 10
2.1.1 Generating the mesh hierarchy . . . . . 11
2.1.2 Multilevel simulations on a given hierarchy . . . . . 13
2.2 Fully stochastic time stepping . . . . . 14
2.3 Algorithm Listings . . . . . 17

**3 Numerical Experiments** **20**
3.1 A Linear SDE . . . . . 21
3.2 Drift singularity, linear SDE . . . . . 24
3.3 Stopped diffusion . . . . . 25

**4 Theoretical results** **32**
4.1 Single level results . . . . . 34
4.2 multilevel results . . . . . 41

**5 Conclusions** **53**

**A Theorems** **54**

# 1 Introduction

This work develops multilevel versions of adaptive algorithms for weak approximation of Itô stochastic differential equations (SDEs)

$$dX(t) = a(t, X(t))dt + b(t, X(t))dW(t), \qquad 0 < t < T, \tag{1.1}$$

where $X(t;\omega)$ is a stochastic process in $\mathbb{R}^d$, with randomness generated by a $k$-dimensional Wiener process with independent components, $W(t;\omega)$; see [15], [23]. The functions $a(t,x) \in \mathbb{R}^d$ and $b(t,x) \in \mathbb{R}^{d\times k}$ are given drift and diffusion fluxes.

Our goal is to, for any given sufficiently well behaved function $g : \mathbb{R}^d \to \mathbb{R}$, approximate the expected value $\mathrm{E}[g(X(T))]$ by adaptive multilevel Monte Carlo (MLMC) methods. A typical example of such an expected value is to compute

option prices in mathematical finance; see [14] and [9]. Other related models based on stochastic dynamics are used for example in molecular dynamics simulations at constant temperature, for stochastic climate prediction, and for wave propagation in random media; cf. [2], [18], and [1].

The computational complexity of a Monte Carlo method is determined by the number of generated samples approximating $g(X(T))$ and their average cost. When a standard Monte Carlo method based on a uniform time stepping scheme of weak order 1 is used to compute $\mathrm{E}[g(X(T))]$ to an accuracy TOL with high probability, the cost is asymptotically proportional to $\mathrm{TOL}^{-3}$, provided that the functions $a$, $b$, and $g$ are sufficiently regular. A Monte Carlo method can not do better than a cost proportional to $\mathrm{TOL}^{-2}$, since this is the cost if each sample of $g(X(T))$ can be generated exactly at a unit cost. The goal of this work is to combine two techniques for improving the standard Monte Carlo: one technique is to use adaptive time stepping which retains the complexity $\mathcal{O}(\mathrm{TOL}^{-3})$ for a wider set of problems than uniform time stepping, and which can reduce the proportionality constant for other problems with widely varying scales; the other is the MLMC method, which in many cases can reduce the complexity to nearly the optimal $\mathcal{O}(\mathrm{TOL}^{-2})$ when based on the Euler–Maruyama time stepping scheme, and which can achieve the optimal rate using the Milstein time stepping scheme.

In the context of weak approximation of SDEs, the MLMC method based on uniform time stepping was introduced by Giles in [7]. A similar MLMC idea has been used before in the different context of parametric integration; [11, 12]. In [7], Giles developed a clever control variate type variance reduction technique for a numerical method, denoted here by $\overline{X}$, that approximates the solution of the SDE (1.1). The key to this variance reduction, which is an extension of a two-level control variate technique in [16], is to compute approximate solutions, $\overline{X}_\ell$, on hierarchies of uniform time meshes with size

$$\Delta t_\ell = C^{-\ell}\Delta t_0, \qquad C \in \{2,3,\ldots\} \quad \text{and} \quad \ell \in \{0,1,\ldots,L\}, \tag{1.2}$$

thereby generating sets of realizations on different mesh levels. After computing numerical approximations on a mesh hierarchy, the expected value $\mathrm{E}[g(X(T))]$ is approximated by the multilevel Monte Carlo estimator

$$\begin{aligned}\mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big); M_0\big) = &\sum_{i=1}^{M_0} \frac{g(\overline{X}_0(T;\omega_{i,0}))}{M_0} \\ &+ \sum_{\ell=1}^{L}\sum_{i=1}^{M_\ell} \frac{g(\overline{X}_\ell(T;\omega_{i,\ell})) - g(\overline{X}_{\ell-1}(T;\omega_{i,\ell}))}{M_\ell}.\end{aligned} \tag{1.3}$$

This estimator combines $L+1$ sample averages, based on mutually independent sample sets on the respective meshes, each with $M_\ell$ independent samples. The number of samples $\{M_\ell\}_{\ell=1}^{L}$ have a given relation to the number of samples on the coarsest mesh, $M_0$, which is the only free parameter in (1.3), for a fixed number of levels. To reduce the variance in the above estimator, the

realization pairs $\overline{X}_\ell(T;\omega_{i,\ell})$ and $\overline{X}_{\ell-1}(T;\omega_{i,\ell})$ of the summands $g(\overline{X}_\ell(T;\omega_{i,\ell}))-g(\overline{X}_{\ell-1}(T;\omega_{i,\ell}))$ for each level $\ell>0$ are generated by the same Brownian path, $W(t;\omega_{i,\ell})$, but they are realized on different temporal grids with uniform time steps, $\Delta t_\ell$ and $\Delta t_{\ell-1}$, respectively. The efficiency of this computation relies on an a priori known order of strong convergence for the numerical method employed on each level of the hierarchy.

Let TOL $>0$ be a desired accuracy in the approximation of $\mathrm{E}[g(X(T))]$. The main result of Giles' work [7] is that the computational cost needed to achieve the Mean Square Error (MSE)

$$\mathrm{E}\left[\left(\mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big);M_0\big)-\mathrm{E}[g(X(T))]\right)^2\right]=\mathcal{O}\left(\mathrm{TOL}^2\right), \tag{1.4}$$

when using the Forward Euler method to create the approximate realizations $\overline{X}_\ell(T;\omega)$, can be reduced from $\mathcal{O}\left(\mathrm{TOL}^{-3}\right)$ with the standard Monte Carlo method to

$$\mathcal{O}\left((\mathrm{TOL}^{-1}\log(\mathrm{TOL}^{-1}))^2\right)$$

with Giles' MLMC method. Furthermore, whenever the function $g$ is Lipschitz and for scalar Itô stochastic differential equations, the computational cost can be further reduced to $\mathcal{O}\left(\mathrm{TOL}^{-2}\right)$ using the first order strong convergence Milstein method. In addition, the work [6] shows how to apply the Milstein method for several scalar SDE cases where the Lipschitz condition is not fulfilled and still obtain the cost $\mathcal{O}\left(\mathrm{TOL}^{-2}\right)$.

In this work we use the Euler–Maruyama method with non uniform time steps. Let $0=t_0<t_1<\cdots<t_N=T$ denote a given time discretization, without reference to its place in the hierarchies, and $\{0=W(t_0;\omega),W(t_1;\omega),\ldots,W(t_N;\omega)\}$ denote a generated sample of the Wiener process on that discretization. Then the Euler–Maruyama approximation to the true solution of (1.1) is given by the scheme

$$\begin{aligned}\overline{X}(t_0;\omega)&=X(0),\\ \overline{X}(t_{n+1};\omega)&=a(\overline{X}(t_n;\omega),t_n)\Delta t_n+b(\overline{X}(t_n;\omega),t_n)\Delta W(t_n;\omega),\quad n\geq 0,\end{aligned} \tag{1.5}$$

where $\Delta t_n=t_{n+1}-t_n$ and $\Delta W(t_n;\omega)=W(t_{n+1};\omega)-W(t_n;\omega)$ are the time steps and Wiener increments, respectively.

The contribution of the present paper to the MLMC method is the development and analysis of two novel algorithms with adaptive, non uniform time steps. One of the algorithms uses adaptive mesh refinements to stochastically create a path dependent mesh for each realization; the other algorithm constructs the meshes adaptively based on sample averaged error densities and then uses the same mesh hierarchy for all realizations. We refer to the previous algorithm as the *stochastic* time step algorithm and the latter as the *deterministic* time step algorithm. The construction and analysis of the adaptive algorithms is inspired by the work on single level adaptive algorithms for weak approximation of ordinary stochastic differential equations [19], and uses the adjoint weighted global error estimates first derived in [24]. The goal of these

adaptive algorithms is to choose the time steps and the number of realizations such that the event

$$\left|\mathcal{A}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big); M_0\big) - \mathrm{E}[g(X(T))]\right| \leq \mathrm{TOL}, \tag{1.6}$$

holds with probability close to one. We measure computational complexity as the work needed to meet a given accuracy. On some problems where the computational complexity of methods based on uniform time steps deteriorates due to lacking regularity, these adaptive mesh refinement algorithms can regain the same rate of convergence that uniform methods have on regular problems; see [20].

It should be noted that in the setting of adaptive mesh refinement there is no given notion of mesh size, so a hierarchy of meshes can no longer be described as in the constant time step case (1.2). Instead, we generate a hierarchy of meshes by successively increasing the accuracy in our computations, introducing the time discretization error tolerance levels[1]

$$\mathrm{TOL}_{\mathrm{T},\ell} = 2^{\ell-L}\mathrm{TOL}_{\mathrm{T}}, \qquad \text{for} \quad \ell \in \{0, 1, \ldots, L\}, \tag{1.7}$$

and (by adaptive refinements based on error indicators) determining the corresponding meshes so that for each level $\ell \in \{0, 1, \ldots, L\}$,

$$\left|\mathrm{E}[g(X(T))] - \mathrm{E}\left[g(\overline{X}_\ell(T))\right]\right| \lesssim \mathrm{TOL}_{\mathrm{T},\ell}.$$

In Section 4, we prove that also for the adaptive algorithms the computational cost for obtaining the error estimate (1.4), with probability close to one, is close to $\mathcal{O}\left(\mathrm{TOL}^{-2}\log(\mathrm{TOL}^{-1})^2\right)$. More precisely, there are two main results on efficiency and accuracy described in Section 4. Regarding accuracy with probability close to one, Theorem 2 states that the approximation errors in (1.2) are asymptotically bounded by the specified error tolerance times a problem independent factor as the tolerance parameter tends to zero. For the efficiency, Theorem 3 states that, depending on technical assumptions on the error density of the adaptive algorithm, the complexity is close to or equal to the standard complexity in the setting of uniform time steps, with explicitly given constants. A completely analogous result holds for the adaptive algorithm with deterministic steps and it is not included here for the sake of brevity; confer [19].

This work also includes three numerical examples, the most relevant ones being one with a drift singularity and one stopped diffusion. In the singularity example multilevel Monte Carlo based on adaptive time steps requires a computational work $\mathcal{O}\left(\mathrm{TOL}^{-2}\log(\mathrm{TOL}^{-1})^2\right)$ and in the stopped diffusion example $\mathcal{O}\left(\mathrm{TOL}^{-2.1}\log(\mathrm{TOL}^{-1})^2\right)$. For both of these examples the observed complexity is close to the optimal, and more efficient than the single level version of the adaptive algorithm.

The rest of this paper is organized as follows: Section 1.1 introduces the notion of error density and error indicators, and recalls useful results for single

[1]For error control, the tolerance is split into a statistical error tolerance and a time discretization error tolerance; $\mathrm{TOL} = \mathrm{TOL}_{\mathrm{S}} + \mathrm{TOL}_{\mathrm{T}}$, cf. Section 2.

level adaptive forward Euler Monte Carlo methods. Section 2 describes the new adaptive multilevel Monte Carlo algorithms. Section 3 presents results from numerical experiments. Finally, Section 4 proves results on accuracy, stopping, and efficiency for a simplified version of the adaptive multilevel algorithms.

## 1.1 A single level posteriori error expansion

Here we recall previous single level results that are useful for the multilevel analysis in Section 4. In particular, we recall adjoint based error expansions with computable leading order term. Assume that the process $X$ satisfies (1.1) and its approximation, $\overline{X}$, is given by (1.5); then the error expansions in Theorem 1.2 and 2.2 of [24] have the form

$$\mathrm{E}\big[g(X(T)) - g(\overline{X}(T))\big] \quad = \quad \mathrm{E}\left[\sum_{n=1}^{N} \rho_n \Delta t_n^2\right] + \text{ higher order terms, } (1.8)$$

where $\rho_n \Delta t_n^2$ are computable error indicators, that is they provide information for further improvement of the time mesh and $\rho_n$ measures the density of the global error in (1.8). A typical adaptive algorithm does two things iteratively:

1. if the error indicators satisfy an accuracy condition then it stops; otherwise
2. the algorithm chooses where to refine the mesh based on the error indicators and then makes an iterative step to 1.

In addition to estimating the global error $\mathrm{E}\big[g(X(T)) - g(\overline{X}(T))\big]$ in the sense of equation (1.8), the error indicators $\rho_n \Delta t_n^2$ also give simple information on where to refine to reach an optimal mesh, based on the almost sure convergence of the density $\rho_n$ as we refine the discretization, see Section 4 in [20].

In the remaining part of this section we state in Theorem 1 a single level error expansion from [24], which can be used with either stochastic or deterministic time steps.

Given an initial time discretization $\Delta t[0](t)$ and, for the stochastic time steps algorithm, refining until[2]

$$|\rho(t,\omega)|\big(\Delta t(t)\big)^2 < \text{constant}, \tag{1.9}$$

we construct a partition $\Delta t(t)$ by repeated halving of intervals so that it satisfies

$$\Delta t(t) = \Delta t[0](t)/2^n \quad \text{for some natural number } n = n(t,\omega).$$

The criterion (1.9) uses an approximate error density function $\rho$, satisfying for $t \in [0,T]$ and all outcomes $\omega$ the uniform upper and lower bounds

$$\rho_{low}(\mathrm{TOL}_\mathrm{T}) \le |\rho(t,\omega)| \le \rho_{up}(\mathrm{TOL}_\mathrm{T}). \tag{1.10}$$

[2]The precise expressions including the constants are given in (2.7) and (2.21) below.

The positive functions $\rho_{low}$ and $\rho_{up}$ are chosen so that $\rho_{up}(\mathrm{TOL_T}) \to +\infty$ as $\mathrm{TOL} \downarrow 0$ while $\rho_{low}(\mathrm{TOL_T}) \to 0$ such that $\mathrm{TOL_T}/\rho_{low}(\mathrm{TOL_T}) \to 0$. We further make the assumption that for all $s, t \in [0, T]$ the sensitivity of the error density to values of the Wiener process can be bounded,

$$|\partial_{W(t)}\rho(s, \omega)| \leq D\rho_{up}(\mathrm{TOL_T}), \tag{1.11}$$

for some positive function $D\rho_{up}$ such that $D\rho_{up}(\mathrm{TOL_T}) \to +\infty$ as $\mathrm{TOL_T} \downarrow 0$. For each realization successive subdivisions of the steps yield the largest time steps satisfying (1.9). The corresponding stochastic increments $\Delta W$ will have the correct distribution, with the necessary independence, if the increments $\Delta W$ related to the new steps are generated by Brownian bridges [15]; that is the time steps are generated by conditional expected values of the Wiener process.

We begin now by stating in the next lemma the regularity conditions to be used in the analysis of the adaptive multilevel algorithms; the lemma corresponds to Lemma 2.1 in [24], while this formulation is given[3] in Lemma 2.1 in [21].

**Lemma 1** (Regularity)**.** (a) *Assume that the following regularity conditions hold:*

(1) *The functions $a(t, x)$ and $b(t, x)$ are continuous in $(t, x)$ and are twice continuously differentiable with respect to $x$.*

(2) *The partial derivatives of first and second order with respect to $x$ of the functions $a$ and $b$ are uniformly bounded.*

(3) *The function $g$ is twice continuously differentiable, and together with its partial derivatives of first and second order it is uniformly bounded.*

*Then the* cost to go *function, defined by*

$$u(t, x) = E\big[g(X(T)) \mid X(t) = x\big], \tag{1.12}$$

*satisfies the Kolmogorov equation*

$$\partial_t u(t, x) + a_k \partial_k u(t, x) + d_{kn} \partial_{kn} u(t, x) = 0, \qquad u(T, \cdot) = g, \tag{1.13}$$

*where we have used Einstein summation convention*[4]*, and where $d_{kn} = \frac{1}{2} b_k^l b_n^l$.*

(b) *Furthermore, if the following regularity conditions are satisfied:*

(1) *The functions $\partial_\beta a(t, \cdot)$ and $\partial_\beta b(t, \cdot)$ are bounded uniformly in $t$ for multi-indices $\beta$ with $1 \leq |\beta| \leq 8$;*

(2) *The functions $a(\cdot, x)$, $b(\cdot, x)$ have continuous and uniformly bounded first order time derivatives;*

[3] Including a jump term omitted here.

[4] When an index variable appears twice in a single term this means that a summation over all possible values of the index takes place; for example $a_k \partial_k u(t, x) = \sum_{k=1}^{d} a_k \partial_k u(t, x)$, where $d$ is the space dimension of the SDE ($a, x \in \mathbb{R}^d$).

(3) *The function $g$ has spatial derivatives $\partial_\beta g$, with polynomial growth for $|\beta| \leq 8$;*

*then the function $u$ has continuous partial derivatives with respect to $x$ up to the order 8, satisfying the following polynomial growth condition: for all $i \in \{0,1,2\}$ and $\alpha \in \mathbb{N}^d$ with $i + |\alpha| \leq 8$ there exists $p_{\alpha,i} \in \mathbb{N}$ and $C_{\alpha,i} > 0$ such that*

$$\max_{0\leq t\leq T} \left| \partial_t^i \partial_\alpha u(t,x) \right| \leq C_{\alpha,i} \left(1 + |x|^{p_{\alpha,i}}\right) \quad \forall\, x \in \mathbb{R}^d.$$

In what follows, Lemma 2 and Theorem 1 show that although the steps adaptively generated to satisfy (1.9)–(1.11) are not adapted to the natural Wiener filtration, the method indeed converges to the correct limit, which is the same as the limit of the forward Euler method with adapted time steps. Lemma 2 and Theorem 1 correspond to Lemma 3.1 and Theorem 3.3 in [24].

**Lemma 2** (Strong Convergence)**.** *For $X$ the solution of (1.1) suppose that $a$, $b$, and $g$ satisfy the assumptions in Lemma 1, that $\overline{X}$ is constructed by the forward Euler method, based on the stochastic time stepping algorithm defined in Section 2, with step sizes $\Delta t_n$ satisfying (1.9)–(1.11), and that their corresponding $\Delta W_n$ are generated by Brownian bridges. Then*

$$\sup_{0\leq t\leq T} \mathrm{E}\left[\left|X(t) - \overline{X}(t)\right|^2\right] = \mathcal{O}\left(\Delta t_{\sup}\right) = \mathcal{O}\left(\frac{\mathrm{TOL}_\mathrm{T}}{\rho_{low}(\mathrm{TOL}_\mathrm{T})}\right) \to 0 \tag{1.14}$$

*as $\mathrm{TOL}_\mathrm{T} \downarrow 0$, where $\Delta t_{\sup} \equiv \sup_{n,\omega} \Delta t_n(\omega)$.*

In Theorem 1 and the rest of this work, we will use Einstein summation convention with respect to functional and spatial indices, but not with respect to the temporal one (usually denoted $t_n$).

**Theorem 1** (Single level stochastic time steps error expansion)**.** *Given the assumptions in Lemma 2 and a deterministic initial value $X(0)$, the time discretization error in (1.8) has the following expansion, based on both the drift and diffusion fluxes and the discrete dual functions $\varphi$, $\varphi'$, and $\varphi''$ given in (1.17)–(1.22), with computable leading order terms:*

$$\begin{aligned}\mathrm{E}[g(X(T))] - \mathrm{E}\left[g(\overline{X}(T))\right] &= \mathrm{E}\left[\sum_{n=0}^{N-1} \tilde{\rho}(t_n,\omega)(\Delta t_n)^2\right] \\ &\quad + \mathcal{O}\left(\left(\frac{\mathrm{TOL}_\mathrm{T}}{\rho_{low}(\mathrm{TOL}_\mathrm{T})}\right)^{1/2}\left(\frac{\rho_{up}(\mathrm{TOL}_\mathrm{T})}{\rho_{low}(\mathrm{TOL}_\mathrm{T})}\right)^{\epsilon}\right) \mathrm{E}\left[\sum_{n=0}^{N-1} (\Delta t_n)^2\right],\end{aligned} \tag{1.15}$$

*for any $\epsilon > 0$ and where*

$$\begin{aligned}\tilde{\rho}(t_n,\omega) \equiv \frac{1}{2}\Big(&\left(\partial_t a_k + \partial_j a_k a_j + \partial_{ij} a_k d_{ij}\right)\varphi_k(t_{n+1}) \\ &+ \left(\partial_t d_{km} + \partial_j d_{km} a_j + \partial_{ij} d_{km} d_{ij} + 2\partial_j a_k d_{jm}\right)\varphi'_{km}(t_{n+1}) \\ &+ \left(2\partial_j d_{km} d_{jr}\right)\varphi''_{kmr}(t_{n+1})\Big)\end{aligned} \tag{1.16}$$

*and the terms in the sum of* (1.16) *are evaluated at the a posteriori known points* $(t_n, \overline{X}(t_n))$*, i.e.,*

$$\partial_\alpha a \equiv \partial_\alpha a(t_n, \overline{X}(t_n)), \quad \partial_\alpha b \equiv \partial_\alpha b(t_n, \overline{X}(t_n)), \quad \partial_\alpha d \equiv \partial_\alpha d(t_n, \overline{X}(t_n)).$$

*Here* $\varphi \in \mathbb{R}^d$ *is the solution of the discrete dual backward problem*

$$\begin{array}{l} \varphi_i(t_n) = \partial_i c_j(t_n, \overline{X}(t_n))\varphi_j(t_{n+1}), \quad t_n < T, \\ \varphi_i(T) = \partial_i g(\overline{X}(T)), \end{array} \tag{1.17}$$

*with*

$$c_i(t_n, x) \equiv x_i + \Delta t_n a_i(t_n, x) + \Delta W_n^\ell b_i^\ell(t_n, x) \tag{1.18}$$

*and its first and second variation*

$$\varphi'_{ij} \equiv \partial_{x_j(t_n)}\varphi_i(t_n) \equiv \frac{\partial \varphi_i(t_n; \overline{X}(t_n) = x)}{\partial x_j}, \tag{1.19}$$

$$\varphi''_{ikm}(t_n) \equiv \partial_{x_m(t_n)}\varphi'_{ik}(t_n) \equiv \frac{\partial \varphi'_{ik}(t_n; \overline{X}(t_n) = x)}{\partial x_m}, \tag{1.20}$$

*which satisfy*

$$\begin{array}{ll} \varphi'_{ik}(t_n) & = \partial_i c_j(t_n, \overline{X}(t_n))\partial_k c_p(t_n, \overline{X}(t_n))\varphi'_{jp}(t_{n+1}) \\ & \quad +\partial_{ik} c_j(t_n, \overline{X}(t_n))\varphi_j(t_{n+1}), \quad t_n < T, \\ \varphi'_{ik}(T) & = \partial_{ik} g(\overline{X}(T)), \end{array} \tag{1.21}$$

*and*

$$\begin{array}{ll} \varphi''_{ikm}(t_n) = & \partial_i c_j(t_n, \overline{X}(t_n))\partial_k c_p(t_n, \overline{X}(t_n))\partial_m c_r(t_n, \overline{X}(t_n))\varphi''_{jpr}(t_{n+1}) \\ & +\partial_{im} c_j(t_n, \overline{X}(t_n))\partial_k c_p(t_n, \overline{X}(t_n))\varphi'_{jp}(t_{n+1}) \\ & +\partial_i c_j(t_n, \overline{X}(t_n))\partial_{km} c_p(t_n, \overline{X}(t_n))\varphi'_{jp}(t_{n+1}) \\ & +\partial_{ik} c_j(t_n, \overline{X}(t_n))\partial_m c_p(t_n, \overline{X}(t_n))\varphi'_{jp}(t_{n+1}) \\ & +\partial_{ikm} c_j(t_n, \overline{X}(t_n))\varphi_j(t_{n+1}), \quad t_n < T, \\ \varphi''_{ikm}(T) = & \partial_{ikm} g(\overline{X}(T)), \end{array} \tag{1.22}$$

*respectively.*

The previous result can also be directly applied to the particular case of deterministic time steps.Observe that the constant in $\mathcal{O}$ that appears in (1.15) may not be uniform with respect to the value $\epsilon$. Thus, in practice one chooses $\epsilon = \epsilon(\mathrm{TOL})$ to minimise the contribution of the remainder term to the error expansion (1.15).

Let us now discuss how to modify the error density $\tilde{\rho}(t_n, \omega)$ in (1.16) to satisfy the bounds (1.10) and at the same time guarantee that $\Delta t_{\sup} \to 0$ as $\mathrm{TOL}_\mathrm{T} \downarrow 0$; see Lemma 2.

Consider, for $t \in [t_n, t_{n+1})$ and $n = 1, \ldots, N$, the piecewise constant function

$$\rho(t) \equiv \mathrm{sign}(\tilde{\rho}(t_n)) \min\big(\max(|\tilde{\rho}(t_n)|, \rho_{low}(\mathrm{TOL}_\mathrm{T})), \rho_{up}(\mathrm{TOL}_\mathrm{T})\big), \tag{1.23}$$

where

$$\begin{array}{l}\rho_{low}(\mathrm{TOL_T}) = \mathrm{TOL_T^{\bar{\gamma}}}, \quad 0 < \bar{\gamma} < \frac{\alpha}{\alpha+2}, \quad 0 < \alpha < \frac{1}{2}, \\ \rho_{up}(\mathrm{TOL_T}) = \mathrm{TOL_T^{-r}}, \quad r > 0,\end{array} \tag{1.24}$$

and with the standard notation for the function sign, that is $\mathrm{sign}(x) = 1$ for $x \geq 0$ and $-1$ for $x < 0$. The function $\rho$ defined by (1.23) measures the density of the time discretisation error; the stochastic time stepping algorithm uses it in (2.20) and (2.21) to guide the mesh refinements, while the deterministic (path independent) time stepping algorithm uses a sample average of it for the same purpose; see (2.6) and (2.7). From now on, with a slight abuse of notation, $\rho(t_n) = \rho_n$ denotes the modified density (1.23).

Following the error expansion in Theorem 1, the time discretization error is approximated by

$$|\mathcal{E}_T| = |\mathrm{E}\left[g(X(T)) - g(\overline{X}(T))\right]| \lesssim \mathrm{E}\left[\sum_{n=1}^{N} r(n)\right] \tag{1.25}$$

using the error indicator, $r(n)$, defined by

$$r(n) \equiv |\rho(t_n)|\Delta t_n^2 \tag{1.26}$$

with the modified error density defined by (1.23). According to Corollary 4.3 and Theorem 4.5 in [19], we have the almost sure convergence of the error density to a limit density denoted by $\hat{\rho}$, $\rho \to \hat{\rho}$ as $\mathrm{TOL_T} \downarrow 0$.

# 2 Adaptive Algorithms and Multilevel Variance Reduction

In this section we will describe two versions of the adaptive MLMC algorithm suitable for different problem settings. The first algorithm version which we present in Section 2.1 is designed for problems with deterministic time dependence on the drift and the diffusion. This algorithm version constructs a mesh hierarchy by adaptive refinements based on comparatively small sample sets and then performs a greater number of realizations on the constructed meshes to control the statistical error. The second algorithm version which we present in Section 2.2 is designed for problems where stochastic effects motivate the adaptive refinements. For this algorithm version the computational meshes are constructed by adaptive refinements for each individual realization of the underlying Wiener process.

## 2.1 Path independent time stepping

We recall that for a given SDE (1.1), function $g : \mathbb{R}^d \to \mathbb{R}$, and tolerance $\mathrm{TOL} > 0$, our goal is to construct an MLMC algorithm for which the event

$$\left|\mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big); M_0\big) - \mathrm{E}[g(X(T))]\right| \leq \mathrm{TOL},$$

holds with probability close to one for the MLMC estimator $\mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big); M_0\big)$ defined by (1.3). We approach this goal by splitting the above approximation error as follows

$$\begin{aligned}&\big|\mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big); M_0\big) - \mathrm{E}[g(X_T)]\big| \\ &\quad\leq \underbrace{\big|\mathrm{E}\big[g(\overline{X}_L(T)) - g(X(T))\big]\big|}_{=:\mathcal{E}_T} + \underbrace{\big|\mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big); M_0\big) - \mathrm{E}\big[g(\overline{X}_L(T))\big]\big|}_{=:\mathcal{E}_S},\end{aligned}$$

and control the total error by requiring that the time discretization error fulfills $\mathcal{E}_T \lesssim \mathrm{TOL}_\mathrm{T}$ and the statistical error fulfills $\mathcal{E}_S \leq \mathrm{TOL}_\mathrm{S}$, where the tolerance also have been split into a time discretization error tolerance and a statistical error tolerance,

$$\mathrm{TOL} = \mathrm{TOL}_\mathrm{T} + \mathrm{TOL}_\mathrm{S}.$$

The computations then naturally divides into two phases. The first phase, consisting of Algorithm 1 and Algorithm 2, constructs a hierarchy of grids to control the time discretization error $\mathcal{E}_T$. The second phase, consisting of Algorithm 3 and Algorithm 4, computes a sufficiently large number of Euler–Maruyama realizations (1.5) on the constructed hierarchy of grids to ensure that $\mathcal{E}_S \leq \mathrm{TOL}_\mathrm{S}$, with probability close to one.

### 2.1.1 Generating the mesh hierarchy

We start with generating a hierarchy of meshes $\{\Delta t_\ell\}_{\ell=0}^L$ for numerical approximation of the SDE (1.1). The meshes are sequentially and adaptively refined from a given initial mesh $\Delta t_{-1}$ such that $\Delta t_{\ell-1} \subset \Delta t_\ell$ for all $\ell \in \{0, 1, \ldots, L\}$. On level $\ell$ the grid is constructed with the aim that the time discretization error in the approximation of $\mathrm{E}\big[g(\overline{X}_\ell(T))\big]$ fulfills

$$\Big|\mathrm{E}\big[g(\overline{X}_\ell(T)) - g(X(T))\big]\Big| < 2^{L-\ell}\mathrm{TOL}_\mathrm{T} =: \mathrm{TOL}_{\mathrm{T},\ell}, \tag{2.1}$$

where $\overline{X}_\ell(T)$ denotes an Euler–Maruyama approximation of the SDE (1.1) on the grid $\Delta t_\ell$. The number of grid levels $L$ is chosen so that the largest tolerance

$$\mathrm{TOL}_{\mathrm{T},0} = 2^L\mathrm{TOL}_\mathrm{T}, \tag{2.2}$$

is much larger than $\mathrm{TOL}_\mathrm{T}$ and results in quite coarse meshes on the lower levels $\Delta t_0, \Delta t_1, \ldots$. To be more precise, with a rough estimate of the magnitude of $\mathrm{E}[g(X(T))]$ taken into account we prescribe an upper bound $\mathrm{TOL}_{\mathrm{T,Max}}$ for $\mathrm{TOL}_{\mathrm{T},0}$ and determine $L$ by the equation

$$L = \lfloor \log_2(\mathrm{TOL}_{\mathrm{T,Max}}/\mathrm{TOL}_\mathrm{T})\rfloor. \tag{2.3}$$

The inputs in Algorithm 1 are: initial mesh $\Delta t_{-1}$, initial number of sample realizations $M_{-1}$, time discretization error tolerance $\mathrm{TOL}_\mathrm{T}$, grid levels $L$, initial estimate of the number of time steps on the accepted coarse mesh $\overline{\mathcal{N}}_0$, and the three parameters $C_\mathrm{R}$, $C_\mathrm{S}$, and $R$ which are all used in the refinement and

stopping conditions (2.7), (2.6), and (2.10), respectively. We choose the initial estimated number of time steps $\overline{\mathcal{N}}_0$ as a small integer no smaller than the number of steps in $\Delta t_{-1}$. For smaller tolerances, $\ell > 0$, the initial guess is calculated according to $\overline{\mathcal{N}}_{\ell,\text{init}} = 2\overline{\mathcal{N}}_{\ell-1,\text{accepted}}$.

The grid refinement algorithm uses sample averages and sample variances of computed error indicators (1.23). With this in mind, we introduce some notation. Consider a set of $M$ independent, identically distributed samples from the probability domain, and for such a sample set, introduce the sample average operator

$$\mathcal{A}(f;M) := \frac{1}{M}\sum_{i=1}^{M} f(\omega_i) \tag{2.4}$$

and, similarly, define the sample variance operator

$$\mathcal{V}(f;M) := \frac{1}{M-1}\sum_{i=1}^{M}\Big(f(\omega) - \mathcal{A}(f;M)\Big)^2. \tag{2.5}$$

When refining the grid $\Delta t_\ell$. the sampled error indicators $r_\ell(n)$, as defined in equation (1.26), are computed for all the time steps of the grid. Let $N_\ell$ denote the number of time steps and $\overline{\mathcal{N}}_\ell$ be an estimate of $N_\ell$ as described above. Then, if the stopping condition

$$\max_{1\le n\le N_\ell} \mathcal{A}(r_\ell(n);M_\ell) < C_{\text{S}}\frac{\text{TOL}_\ell}{\overline{\mathcal{N}}_\ell}, \tag{2.6}$$

is fulfilled, the grid is accepted and the refinements stop; otherwise the $n$-th time step is refined by splitting it into two equal parts if

$$\mathcal{A}(r_\ell(n);M_\ell) \ge C_{\text{R}}\frac{\text{TOL}_\ell}{\overline{\mathcal{N}}_\ell}. \tag{2.7}$$

Normally, the value for $C_{\text{R}}$ would be around 2, and one must take $C_{\text{S}} > C_{\text{R}}$ following the theory developed in [20, 19].

The adaptive refinements of the computational grid are based on the sample averaged error indicators $\mathcal{A}(r_\ell(n);M_\ell)$. To estimate the mean error indicators $\text{E}[r_\ell(n)]$ with sufficient accuracy, we need a mechanism for deciding how many samples to use in the sample averages. With $\texttt{E}_{\Delta t_\ell}$ denoting the computed estimate of the time discretization error, i.e.,

$$\texttt{E}_{\Delta t_\ell} = \sum_{n=1}^{N_\ell} \mathcal{A}(r_\ell(n);M_\ell)\,, \tag{2.8}$$

a reasonable reliability requirement is

$$\sqrt{\text{Var}(\texttt{E}_{\Delta t_\ell})} < R\,\text{E}[\texttt{E}_{\Delta t_\ell}], \tag{2.9}$$

for some suitably chosen $0 < R < 1$. (In our numerical examples, for instance, we use $R = 0.2$.) The variance of $\texttt{E}_{\Delta t_\ell}$ is however unknown, but the i.i.d.

distribution of the error indicators sampled motivates the approximation

$$\mathrm{Var}(\mathtt{E}_{\Delta t_\ell}) \approx \frac{\mathcal{V}\Big(\sum_{n=1}^{N_\ell} r_\ell(n); M_\ell\Big)}{M_\ell} \quad \text{for } \ell = 0, 1, \ldots, L.$$

We consequently approximate the reliability requirement (2.9) by

$$\sqrt{\frac{\mathcal{V}\Big(\sum_{n=1}^{N_\ell} r_\ell(n); M_\ell\Big)}{M_\ell}} < R\,\mathtt{E}_{\Delta t_\ell}, \quad \text{for } \ell = 0, 1, \ldots, L, \tag{2.10}$$

where the number of sample realizations $M_\ell$ used on level $\ell$ in the grid construction phase is increased by repeated doubling, i.e., $M_{\ell,new} = 2\,M_{\ell,old}$, until inequality (2.10) is satisfied. The initial batch size at each level is set by $M_\ell = M_{\ell-1}$, where we for the moment let $M_{\ell-1}$ denote the stopped number of samples at level $\ell - 1$, and for level $\ell = 0$ it turns out to be sufficient to use initial batch size $M_0 = M_{-1}$ with

$$M_{-1} = \mathrm{const} \cdot \mathrm{TOL}_{\mathrm{T}}^{-1}. \tag{2.11}$$

The adaptive algorithm that generates the above described mesh hierarchy for the deterministic time step version of the MLMC algorithm is presented in Algorithm 1–2 in Section 2.3.

### 2.1.2 Multilevel simulations on a given hierarchy

In the second phase we will describe the algorithms which ensure that our MLMC estimate of $\mathrm{E}\big[g(\overline{X}_L(T))\big]$ fulfills the statistical error bound

$$\mathcal{E}_S = \big|\mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big); M_0\big) - \mathrm{E}\big[g(\overline{X}_L(T))\big]\big| \leq \mathrm{TOL}_{\mathrm{S}}, \tag{2.12}$$

with probability close to one. We recall from (1.3) that the MLMC estimator is defined by

$$\mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big); M_0\big) = \mathcal{A}\big(g(\overline{X}_0(T)); M_0\big) + \sum_{\ell=1}^{L} \mathcal{A}\big(g(\overline{X}_\ell(T)) - g(\overline{X}_{\ell-1}(T)); M_\ell\big), \tag{2.13}$$

where the realization pairs $\overline{X}_\ell(T;\omega_{i,\ell})$ and $\overline{X}_{\ell-1}(T;\omega_{i,\ell})$ of the summands $g(\overline{X}_\ell(T;\omega_{i,\ell})) - g(\overline{X}_{\ell-1}(T;\omega_{i,\ell}))$ for each level $\ell > 0$ are generated by the Euler–Maruyama method (1.5) using the same Brownian path $W(t;\omega_{i,\ell})$ on the respective *different* temporal meshes $\Delta t_\ell$ and $\Delta t_{\ell-1}$ that were computed by Algorithm 1. Furthermore, all Brownian paths $\{W(t;\omega_{i,\ell})\}_{i,\ell}$ are independent, and the number of samples at the coarsest level is set to $M_0 = 2^{L+\lceil C_{\mathcal{ML}} L\rceil+1}$ for a suitable constant $C_{\mathcal{ML}} \in (0,1)$, cf. Remark 3, and the number of samples on higher levels is expressed in terms of $M_0$ by the ratio

$$M_\ell = \frac{M_0}{2^L}\left\lceil 2^L \frac{\rho_{low}(\mathrm{TOL}_{\mathrm{T},0})\mathrm{TOL}_{\mathrm{T},\ell}}{\rho_{low}(\mathrm{TOL}_{\mathrm{T},\ell})\mathrm{TOL}_{\mathrm{T},0}}\right\rceil, \quad \ell = 1, \ldots, L, \tag{2.14}$$

where $\rho_{low}$ is the lower bound for the error density introduced in (1.24) and $\lceil \cdot \rceil$ denotes rounding upwards to the nearest integer. The enforced lower bound for the sample sets $\{M_\ell\}_{\ell=0}^L$ implies that $M_L \to \infty$ as $\mathrm{TOL} \downarrow 0$, and this motivates the approximation of

$$\frac{\mathcal{A}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big); M_0\big) - \mathrm{E}\big[g(\overline{X}_L(T))\big]}{\sqrt{\mathrm{Var}\big(\mathcal{A}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big); M_0\big)\big)}}$$

by a normal distributed random variable; see Lemma 8 in Section 4 for a justification of this approximation for the stochastic time step setting. Relying on this approximation, the statistical error (2.12) will be controlled by bounding the MLMC estimator variance $\sqrt{\mathrm{Var}\big(\mathcal{A}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big); M_0\big)\big)} \leq C_C \mathrm{TOL}_\mathrm{S}$, for a given positive confidence parameter $C_C$. The variance $\mathrm{Var}\big(\mathcal{A}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big); M_0\big)\big)$ is however unknown, so we introduce the following approximation

$$\mathrm{Var}\big(\mathcal{A}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big); M_0\big)\big) \approx \underbrace{\frac{\mathcal{V}\big(g(\overline{X}_0(T)); M_0\big)}{M_0} + \sum_{\ell=1}^{L} \frac{\mathcal{V}\big(g(\overline{X}_\ell(T)) - g(\overline{X}_{\ell-1}(T)); M_\ell\big)}{M_\ell}}_{=:\sigma^2}. \tag{2.15}$$

Our stopping criterion for the Monte Carlo simulations then becomes

$$\sigma < \frac{\mathrm{TOL}_\mathrm{S}}{C_\mathrm{C}}. \tag{2.16}$$

Until this condition is fulfilled, the number of samples are iteratively doubled ($M_0 = 2M_0$) and the number of samples at the levels $\{M_\ell\}_{\ell=1}^L$ are updated according the ratio (2.14), and a new sample estimate $\mathcal{A}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big); M_0\big)$ is generated using the MLMC estimator (2.13). Having determined $M_0$, we lastly generate and return the output MLMC estimate $\mathcal{A}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big); M_0\big)$.

The probability of controlling the statistical error, i.e., fulfilling the event (2.12) depends on the chosen value for the confidence parameter $C_\mathrm{C}$. For example, with $C_\mathrm{C} = 1.65$ the event

$$\Big|\mathcal{A}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big); M_0\big) - \mathrm{E}\big[g(\overline{X}_L(T))\big]\Big| < C_\mathrm{C}\,\sigma,$$

occurs with probability greater than 0.9, asymptotically as $\mathrm{TOL} \downarrow 0$. See Algorithm 3–4 in Section 2.3 for more details on the MLMC algorithms approximating $\mathrm{E}\big[g(\overline{X}_L(T))\big]$ in the path independent time step setting.

## 2.2 Fully stochastic time stepping

In this subsection we describe the MLMC algorithm for approximating $\mathrm{E}[g(X(T))]$ in the setting with adaptive stochastic time steps.

Quite similar to the setting of path independent time steps, the error control of the MLMC estimate $\big|\mathcal{A}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big); M_0\big) - \mathrm{E}[g(X(T))]\big|$ is in this setting

based on constructing numerical realizations $\overline{X}_\ell(t)$ on *stochastic* adaptively refined meshes $\Delta t_\ell$ so that the time discretization errors

$$\left|\mathrm{E}\big[g(\overline{X}_\ell(T)) - g(X(T))\big]\right| \lesssim \mathrm{TOL}_{\mathrm{T},\ell}, \qquad \text{for } \ell = 0, 1, \ldots, L, \tag{2.17}$$

are fulfilled, and by determining the number of samples $M_0$ to ensure that the statistical error

$$\left|\mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big); M_0\big) - \mathrm{E}\big[g(\overline{X}_L(T))\big]\right| \leq \mathrm{TOL}_{\mathrm{S}}, \tag{2.18}$$

is fulfilled, with a given confidence.

The statistical error (2.18) is controlled very similarly as in the setting of path independent time steps:

1. set the initial of samples used in the MLMC estimator (2.13) as $M_0 = 2^{L+\lceil C_{\mathcal{ML}} L\rceil+1}$ with $C_{\mathcal{ML}} \in (0,1)$, cf. Remark 3;

2. configure the number of samples $M_\ell$ on higher levels in terms of $M_0$ by the ratio (2.14);

3. generate realizations $\{\overline{X}_\ell(T)\}$ for the MLMC estimator $\mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big); M_0\big)$ and compute the sample variance $\sigma^2$ as defined in (2.15);

4. If the stopping condition (2.16) is fulfilled, generate a last output MLMC estimate $\mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big); M_0\big)$ and break. Otherwise, set $M_0 = 2M_0$, update the stochastic algorithm parameters estimating the average number of time steps on each grid level,[5] and return to step 2.

For the $\ell$-th sample average summand of $\mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big); M_0\big)$, i.e., $\mathcal{A}\big(g(\overline{X}_0(T); M_0\big)$ if $\ell = 0$ and $\mathcal{A}\big(g(\overline{X}_\ell(T) - g(\overline{X}_{\ell-1}(T)); M_\ell\big)$ if $\ell > 0$, the algorithm generates $M_\ell$ Euler–Maruyama realization pairs[6], $(\overline{X}_{\ell-1}(T), \overline{X}_\ell(T))$ according to (1.5) with the time discretization errors respectively bounded by $\mathrm{TOL}_{\mathrm{T},\ell-1}$ and $\mathrm{TOL}_{\mathrm{T},\ell}$ in the sense (2.17). The realization pairs are constructed by stochastic adaptive refinements of a given initial mesh $\Delta t_{-1}$. The realizations in a realization pair $(\overline{X}_{\ell-1}(T), \overline{X}_\ell(T))$ are respectively generated on the adaptively refined meshes $\Delta t_{\ell-1}$ and $\Delta t_\ell$. These meshes are determined by iteratively refining an initial mesh $\Delta t_{-1}$. First, $\Delta t_{-1}$ is adaptively refined to a mesh $\Delta t_0$ on which $\left|\mathrm{E}\big[g(\overline{X}_0(T)) - g(X(T))\big]\right| \lesssim \mathrm{TOL}_{\mathrm{T},0}$ is fulfilled. Thereafter, $\Delta t_0$ is adaptively refined to a mesh $\Delta t_1$ on which $\left|\mathrm{E}\big[g(\overline{X}_1(T)) - g(X(T))\big]\right| \lesssim \mathrm{TOL}_{\mathrm{T},1}$ is fulfilled. This iterative refinement procedure continues until the mesh $\Delta t_{\ell-2}$ is adaptively refined to generate the first output mesh $\Delta t_{\ell-1}$ and, lastly, $\Delta t_{\ell-1}$ is adaptively refined to generate the second output mesh $\Delta t_\ell$.

The iterative adaptive mesh refinement procedure in Algorithm 5 ensures that a mesh $\Delta t_\ell$ for the fine realization in a pair $(\overline{X}_{\ell-1}(T), \overline{X}_\ell(T))$ is determined in the same way as a mesh $\Delta t_\ell$ for the coarse realization in pair

[5] See Algorithm 6 for details on the parameter update.

[6] Observe that for the level $\ell = 0$ only the realizations of $\overline{X}_0(T)$ are generated.

$(\overline{X}_\ell(T), \overline{X}_{\ell+1}(T))$, and consequently that $\mathrm{E}\big[\overline{X}_\ell(T)\big]$ when computed from the finer realization in a pair $(\overline{X}_{\ell-1}(T), \overline{X}_\ell(T))$ is equal to $\mathrm{E}\big[\overline{X}_\ell(T)\big]$ when computed from the coarse realization in a pair $(\overline{X}_\ell(T), \overline{X}_{\ell+1}(T))$. This property is useful since it implies that the following consistency condition for our MLMC estimator is fulfilled

$$\mathrm{E}\big[\mathcal{A}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big); M_0\big)\big] = \mathrm{E}\big[g(\overline{X}_L(T))\big].$$

Let us next take a closer look at the mesh refinement. Due to the stochastic nature of SDEs, each realization pair $(\overline{X}_{\ell-1}(T), \overline{X}_\ell(T))$ may refine the initial mesh $\Delta t_{-1}$ differently. In particular, meshes corresponding to different realizations on a given level $\ell$ may differ. To describe the mesh refinement, taking this feature into account, we introduce some notation. Let $N_\ell$ and $\overline{\mathcal{N}}_\ell$ denote the number of time steps and the approximate average number of time steps for realizations at level $\ell$, respectively; see Algorithm 6 for details on the approximation technique and its update through the iteration. Further, denote the grid corresponding to one realization at level $\ell$ by

$$\Delta t_\ell = \left[\Delta t_\ell(0), \ldots, \Delta t_\ell(N_\ell - 1)\right], \tag{2.19}$$

and its corresponding Wiener increments by

$$\Delta W_\ell = \left[\Delta W_\ell(0), \ldots, \Delta W_\ell(N_\ell - 1)\right].$$

The refinement condition is based on the error indicator $r_\ell$, defined in (1.26), and uses similar refinements to those in the single level method. The refinements of $\Delta t_\ell$ are stopped when

$$\max_{1 \le n \le N_\ell} r_\ell(n) < C_{\mathrm{S}} \frac{\mathrm{TOL}_{\mathrm{T},\ell}}{\overline{\mathcal{N}}_\ell}. \tag{2.20}$$

As long as inequality (2.20) is violated, the $n^{\text{th}}$ time step of $\Delta t_\ell$ is refined if

$$r_\ell(n) \ge C_{\mathrm{R}} \frac{\mathrm{TOL}_{\mathrm{T},\ell}}{\overline{\mathcal{N}}_\ell}. \tag{2.21}$$

Normally, the value for $C_{\mathrm{R}}$ would be around 2, and $C_{\mathrm{S}} > C_{\mathrm{R}}$ following the theory developed in [20, 19].

A detailed description of the adaptive MLMC algorithm is given in Algorithm 5 with subroutines Algorithm 6–7 in Section 2.3.

The inputs in Algorithm 5 are: $\mathrm{TOL}_{\mathrm{S}}$, $\mathrm{TOL}_{\mathrm{T}}$, initial number of sample realizations $M_0$, $L$, $\Delta t_{-1}$, initial guesses for the mean number of time steps $\{\overline{\mathcal{N}}_{\ell,\mathrm{init}}\}_{\ell=0}^{L}$ of the accepted meshes from the adaptive refinements, and the three parameters $C_{\mathrm{R}}$, $C_{\mathrm{C}}$, and $C_{\mathrm{S}}$ used in the refinement condition (2.21) and stopping conditions (2.16) and (2.20), respectively. In this algorithm the initial estimate of the mean number of time steps are chosen as $\overline{\mathcal{N}}_{\ell,\mathrm{init}} = c\mathrm{TOL}_{\mathrm{T},\ell}{}^{-1}$, for $\ell = 0, \ldots, L$ and a constant $c$ such that $\overline{\mathcal{N}}_{0,\mathrm{init}}$ is a small integer; in the numerical examples in Section 3 the constant was chosen so that $\overline{\mathcal{N}}_{0,\mathrm{init}} \approx 10$.

## 2.3 Algorithm Listings

**Algorithm 1**: Adaptive Generation of a Mesh Hierarchy

**Input** : $\mathrm{TOL}_\mathrm{T}$, $M_{-1}$, $\Delta t_{-1}$, $L$, $\overline{\mathcal{N}}_0$, $C_\mathrm{R}$, $C_\mathrm{S}$, $R$
**Output**: $\{\Delta t_\ell\}_{\ell=0}^{L}$, $M_L$

**for** $\ell = 0, 1, \ldots, L$ **do**
  Set keep_sampling = TRUE, keep_refining = TRUE,
  $\Delta t_\ell = \Delta t_{\ell-1}$, $M_\ell = M_{\ell-1}$, and $\mathrm{TOL}_{\mathrm{T},\ell} = 2^{L-\ell}\mathrm{TOL}_\mathrm{T}$.
  **while** keep_sampling **or** keep_refining **do**
    Set keep_sampling = FALSE, keep_refining = FALSE
    Compute $r_\ell$, $\mathtt{E}_{\Delta t_\ell}$, and $\mathcal{V}\Big(\sum_{n=1}^{N_\ell} r_\ell(n); M_\ell\Big)$ by calling
    Algorithm 2: Euler($M_\ell$, $\Delta t_\ell$)
    **if** $\mathcal{V}\Big(\sum_{n=1}^{N_\ell} r_\ell(n); M_\ell\Big)$ and $\mathtt{E}_{\Delta t}$ violates (2.10) **then**
      Set keep_sampling = TRUE
      Update the number of samples by
      $M_\ell = 2M_\ell$
    **else**
      **if** $r_\ell$ violates (2.6) **then**
        Set keep_refining = TRUE
        Refine $\Delta t_\ell$ by
        **forall** intervals $n = 1, 2, \ldots, N_\ell$ **do**
          **if** $r_\ell(n)$ satisfies (2.7) **then**
            divide the interval $n$ into two equal parts
          **end**
        **end**
        Set $\overline{\mathcal{N}}_\ell = \max\{\overline{\mathcal{N}}_\ell, \mathrm{new}N_\ell\}$.
      **end**
    **end**
  **end**
  Set $\overline{\mathcal{N}}_{\ell+1} = 2\,\overline{\mathcal{N}}_\ell$
**end**

**Algorithm 2**: Euler

**Input** : $M_\ell$, $\Delta t_\ell$
**Output**: $r_\ell$, $\mathtt{E}_{\Delta t_\ell}$, $\mathcal{V}\Big(\sum_{n=1}^{N_\ell} r_\ell(n); M_\ell\Big)$

Compute $M_\ell$ *new* realizations of $\overline{X}_\ell$ on $\Delta t_\ell$ by Euler–Maruyama method (1.5) and use them to compute the error indicators $r_\ell(n)$ on $\Delta t_\ell$ by equation (1.26), $\mathtt{E}_{\Delta t_\ell}$ by equation (2.8), and $\mathcal{V}\Big(\sum_{n=1}^{N_\ell} r_\ell(n); M_\ell\Big)$ by equation (2.5).

**Algorithm 3**: Multilevel Monte Carlo on a Mesh Hierarchy

**Input** : $\mathrm{TOL}_{\mathrm{S}}$, $M_0$, $L$, $\{\Delta t_\ell\}_{\ell=0}^{L}$, $C_{\mathrm{C}}$
**Output**: $\mu = \mathcal{A}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big); M_0\big)$

Set $k = 0$.
**while** $k < 1$ **or** (2.16) **is violated do**
  Set $\mu = 0$ and $\sigma^2 = 0$.
  **for** $\ell = 0, 1, \ldots, L$ **do**
    Set $M_\ell$ as in (2.14)
    **if** $\ell = 0$ **then**
      Call Algorithm 4: Euler$(M_0, \{\Delta t_0\})$.
      Set $\mu = \mu + \mathcal{A}\big(g\left(\overline{X}_0(T)\right); M_\ell\big)$
      and $\sigma^2 = \sigma^2 + \frac{\mathcal{V}\left(g(\overline{X}_0(T)); M_0\right)}{M_0}$ .
    **else**
      Call Algorithm 4: Euler$(M_\ell, \{\Delta t_\ell, \Delta t_{\ell-1}\})$.
      Set $\mu = \mu + \mathcal{A}\big(g\left(\overline{X}_\ell(T)\right) - g\left(\overline{X}_{\ell-1}(T)\right); M_\ell\big)$
      and $\sigma^2 = \sigma^2 + \frac{\mathcal{V}\left(g\left(\overline{X}_\ell(T)\right) - g\left(\overline{X}_{\ell-1}(T)\right); M_\ell\right)}{M_\ell}$ .
    **end**
  **end**
  **if** $\sigma$ violates (2.16) **then**
    Update the number of samples by
    $M_0 = 2M_0$
  **end**
  Increase $k$ by 1
**end**

Generate and return the output $\mu = \mathcal{A}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big); M_0\big)$ according to (2.13).

**Algorithm 4**: Euler

**Input** : $M$, $\{\Delta t_\ell\}_{\ell=l_0,l_1}$
**Output**: $\mathcal{V}\big(g\left(\overline{X}_0(T)\right); M\big)$, $\mathcal{A}\big(g\left(\overline{X}_0(T)\right); M\big)$ if $l_0 = l_1 = 0$ **or** $\mathcal{V}\big(g\left(\overline{X}_{\ell_1}(T)\right) - g\left(\overline{X}_{\ell_0}(T)\right); M\big)$, $\mathcal{A}\big(g\left(\overline{X}_{\ell_1}(T)\right) - g\left(\overline{X}_{\ell_0}(T)\right); M\big)$ if $l_0 \neq l_1$

Simulate $M$ *new* outcomes of the Wiener process $W(t)$ on $\Delta t_{\ell_1} \supseteq \Delta t_{\ell_0}$.
**if** $l_0 = l_1 = 0$ **then**
  Compute the corresponding realizations of $\overline{X}_0$ on $\Delta t_0$ and use them to compute $\mathcal{A}\big(g\left(\overline{X}_0(T)\right); M\big)$ and $\mathcal{V}\big(g\left(\overline{X}_0(T)\right); M\big)$ by (2.4) and (2.5).
**else**
  Compute the corresponding realizations of $\overline{X}_{\ell_1}$ and $\overline{X}_{\ell_0}$ on $\Delta t_{\ell_1}$ and $\Delta t_{\ell_0}$ and use them to compute $\mathcal{A}\big(g\left(\overline{X}_{\ell_1}(T)\right) - g\left(\overline{X}_{\ell_0}(T)\right); M\big)$ and $\mathcal{V}\big(g\left(\overline{X}_{\ell_1}(T)\right) - g\left(\overline{X}_{\ell_0}(T)\right); M\big)$ by (2.4) and (2.5).
**end**

**Algorithm 5**: Multilevel Monte Carlo with stochastic time stepping

**Input** : $\mathrm{TOL}_\mathrm{S}$, $\mathrm{TOL}_\mathrm{T}$, $M_0$, $\Delta t_{-1}$, $L$, $\{\overline{\mathcal{N}}_\ell\}_{\ell=0}^L$, $C_\mathrm{R}$, $C_\mathrm{S}$, $C_\mathrm{C}$
**Output**: $\mu = \mathcal{A}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big); M_0\big)$

Set $k = 0$.
**while** $k < 1$ **or** (2.16) is violated **do**
  Compute $M_0$ new realizations of $g\left(\overline{X}_0(T)\right)$
  and their corresponding number of time steps, $\{N_{0,m}\}_{m=1}^{M_0}$,
  by generating Wiener increments $\{\Delta W_{-1,m}\}_{m=1}^{M_0}$ on the mesh $\Delta t_{-1}$
  (independently for each realization $m$) and calling Algorithm 7:
  **ATSSE**$(\Delta t_{-1}, \Delta W_{-1,m}, \mathrm{TOL}_\mathrm{T} 2^L, \overline{\mathcal{N}}_0)$.
  Set $\mu = \mathcal{A}\big(g\left(\overline{X}_0(T)\right); M_0\big)$ and $\sigma^2 = \frac{\mathcal{V}\left(g(\overline{X}_0(T)); M_0\right)}{M_0}$.
  Compute the average number of time steps $\mathcal{A}(N_0; M_0)$.
  **for** $\ell = 1, \ldots, L$ **do**
    Set $M_\ell$ as in (2.14)
    Compute $M_\ell$ new realizations of $g\left(\overline{X}_{\ell-1}(T)\right)$,
    their corresponding number of time steps, $\{N_{\ell-1,m}\}_{m=1}^{M_\ell}$, and
    Wiener increments, $\{\Delta W_{\ell-1,m}\}_{m=1}^{M_\ell}$, by generating Wiener steps
    $\{\Delta W_{-1,m}\}_{m=1}^{M_0}$ on the mesh $\Delta t_{-1}$ (independently for each
    realization $m$) and using the loop
    **for** $\hat{\ell} = 0, \ldots, \ell - 1$ **do**
      compute $\Delta t_{\hat{\ell},m}$ and $\Delta W_{\hat{\ell},m}$ by calling Algorithm 7:
      **ATSSE**$(\Delta t_{\hat{\ell}-1,m}, \Delta W_{\hat{\ell}-1,m}, \mathrm{TOL}_\mathrm{T} 2^{L-\hat{\ell}}, \overline{\mathcal{N}}_{\hat{\ell}})$.
    **end**
    Compute the corresponding $M_\ell$ realizations of $g\left(\overline{X}_\ell(T)\right)$ and
    their number of time steps, $\{N_{\ell,m}\}_{m=1}^{M_\ell}$, by calling Algorithm 7:
    **ATSSE**$(\Delta t_{\ell-1,m}, \Delta W_{\ell-1,m}, \mathrm{TOL}_\mathrm{T} 2^{L-\ell}, \overline{\mathcal{N}}_\ell)$.
    Set $\mu = \mu + \mathcal{A}\big(g\left(\overline{X}_\ell(T)\right) - g\left(\overline{X}_{\ell-1}(T)\right); M_\ell\big)$ and
    $\sigma^2 = \sigma^2 + \frac{\mathcal{V}\left(g(\overline{X}_\ell(T)) - g(\overline{X}_{\ell-1}(T)); M_\ell\right)}{M_\ell}$.
    Compute the average number of time steps $\mathcal{A}(N_{\ell-1}; M_\ell)$ and
    $\mathcal{A}(N_\ell; M_\ell)$.
  **end**
  **if** $\sigma$ violates (2.16) **then**
    Update the number of samples by
    $M_0 = 2M_0$.
    Update the values of $\{\overline{\mathcal{N}}_\ell\}_{\ell=0}^L$ by calling Algorithm 6:
    **UMNT** $(\{M_\ell\}_{\ell=0}^L,\ \{\mathcal{A}(N_\ell; M_\ell)\}_{\ell=0}^L,\ \{\mathcal{A}(N_{\ell-1}; M_\ell)\}_{\ell=1}^L)$.
  **end**
  Increase $k$ by 1.
**end**

Generate and return the output $\mu = \mathcal{A}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big); M_0\big)$ according to (2.13).

**Algorithm 6**: Update for the mean number of time steps, (**UMNT**)

**Input** : $\{M_\ell\}_{\ell=0}^{L}$, $\{\mathcal{A}(N_\ell; M_\ell)\}_{\ell=0}^{L}$, $\{\mathcal{A}(N_{\ell-1}; M_\ell)\}_{\ell=1}^{L}$
**Output**: $\{\overline{\mathcal{N}}_\ell\}_{\ell=0}^{L}$

**for** $\ell = 0, 1, \ldots, L$ **do**
  **if** $\ell < L$ **then**
    Set $\overline{\mathcal{N}}_\ell = \frac{M_\ell \mathcal{A}(N_\ell; M_\ell) + M_{\ell+1}\mathcal{A}(N_\ell; M_{\ell+1})}{M_\ell + M_{\ell+1}}$.
  **else**
    Set $\overline{\mathcal{N}}_L = \mathcal{A}(N_L; M_L)$.
  **end**
**end**

**Algorithm 7**: Adaptive Time Step Stochastic Euler **(ATSSE)**

**Input** : $\Delta t_{in}, \Delta W_{in}$, TOL, $\overline{\mathcal{N}}_{in}$
**Output**: $\Delta t_{out}, \Delta W_{out}$, $N_{out}$, $g_{out}$
Set $k = 0$, $\Delta t_{[0]} = \Delta t_{in}$, $\Delta W_{[0]} = \Delta W_{in}$, $N_{[0]}$ = number of steps in $\Delta t_{in}$
**while** $k < 1$ **or** $(r_{[k-1]};\ \text{TOL}, \overline{\mathcal{N}}_{in})$ violates (2.20) **do**
  Compute the Euler approximation $\overline{X}_{[k]}$ and the error indicators $r_{[k]}$ on $\Delta t_{[k]}$ with the known Wiener increments $\Delta W_{[k]}$.
  **if** $(r_{[k]};\ \text{TOL}, \overline{\mathcal{N}}_{in})$ violates (2.20) **then**
    Refine the grid $\Delta t_{[k]}$ by
    **forall** intervals $n = 1, 2, \ldots, N_{[k]}$ **do**
      **if** $r_{[k]}(n)$ satisfies (2.21) **then**
        divide the interval $n$ into two equal parts
      **end**
    **end**
    and store the refined grid in $\Delta t_{[k+1]}$.
    Compute $\Delta W_{[k+1]}$ from $\Delta W_{[k]}$ using Brownian bridges on $\Delta t_{[k+1]}$.
    Set $N_{[k+1]}$ = number of steps in $\Delta t_{[k+1]}$.
  **end**
  Increase $k$ by 1.
**end**
Set $\Delta t_{out} = \Delta t_{[k-1]}$, $\Delta W_{out} = \Delta W_{[k-1]}$, $N_{out} = N_{[k-1]}$, $g_{out} = g(\overline{X}_{[k-1]})$.

# 3 Numerical Experiments

This section presents numerical results from implementations of the algorithms of Section 2. We have selected problems to indicate the use of the adaptive methods. Specifically, uniform time steps are suitable for problem 3.1, adaptively refined deterministic time steps are suitable for problem 3.2, and fully stochastic time steps are suitable problem 3.3. In both problems 3.2 and 3.3 the use of the multilevel adaptive algorithms is much more efficient than the use of the corresponding single level versions of the algorithms, which is in turn much more efficient than using a single level uniform time stepping method.

For those problems the complexity is close to that of uniform MLMC, since the observed order of strong convergence remains close to 1/2 even though the order of weak convergence is reduced using uniform time steps. The current adaptive algorithm is not optimized with respect to the strong error, but is subject to ongoing research as an extension of the present adaptive algorithm.

The computations were performed in `Matlab 7` using the built in pseudo random number generator `randn` for simulating sampling from the normal distribution. For the parameter in the refinement criteria (2.7) and (2.21) we used $C_{\mathrm{R}} = 2$ in all the cases, and for the parameter in the stopping criteria (2.6) and (2.20) we used $C_{\mathrm{S}} = 5$ in problems 3.2 and 3.3, and $C_{\mathrm{S}} = 3$ in problem 3.1 where we expect uniform refinements and all error indicators of the same size. In all examples the error tolerance was split equally, $\mathrm{TOL}_{\mathrm{S}} = \mathrm{TOL}_{\mathrm{T}} = \mathrm{TOL}/2$; the proof of Theorem 3 indicates that this choice is not optimal.

## 3.1 A Linear SDE

*Consider first the standard geometric Brownian Motion,*

$$\begin{aligned} dX(t) &= rX(t)dt + \sigma X(t)\, dW(t), \qquad && t \in (0,T), \\ X(0) &= 1, \end{aligned}$$

*using* $r = 1$ *and* $\sigma = 0.5$ *with a final time* $T = 1$ *and* $g(x) = x$.

In this simple example adaptive time stepping is not expected to improve the time discretization error. In fact, the path independent adaptive algorithm produces a hierarchy of uniform grids, and when the fully stochastic adaptive algorithm is applied to this problem all generated meshes are uniform but different realizations of the driving Wiener process may result in different step sizes. The computational cost, measured as the total number of time steps, in all stages in the adaptive refinements, for all realizations of the Euler approximation $\overline{X}$, is shown in Figure 1. For both versions of the algorithm, the computational cost is consistent with, but slightly better than, the main complexity result in Theorem 3 of Section 4, from which we get an upper bound $\mathit{cost} \leq C(\mathrm{TOL}^{-1} \log(\mathrm{TOL}_{\mathrm{T},0}/\mathrm{TOL}))^2$. The work measured this way is very similar in the two versions of the algorithm. However, the version in Section 2.1 is more efficient in this case since it only computes dual solutions in the construction of the mesh hierarchy which is of negligible cost[7], while the version in Section 2.2 computes both primal and dual for every realization. The accuracy of both versions of the algorithm is shown in Figure 2.

The work we measure in Figure 1 is greater than the work analyzed in Section 4, which only counts the number of Euler steps used on the accepted meshes. The comparison made in Table 1 shows the same growth rate as $\mathrm{TOL} \downarrow 0$ when the fully stochastic adaptive algorithm is applied to problem 3.1.

[7] See Figure 3 for problem 3.2.

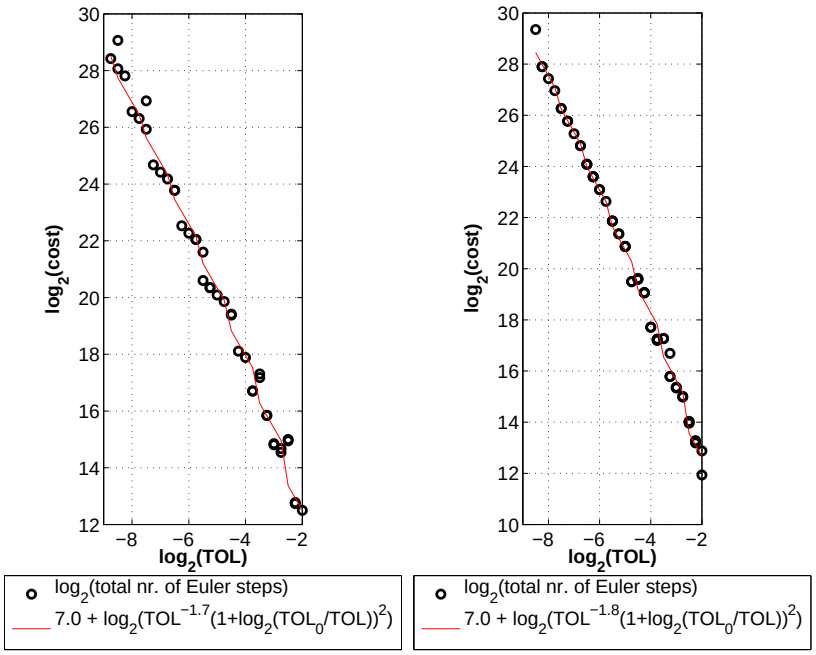


Figure 1: Experimental complexity for both versions of the algorithm applied to the geometrical Brownian motion example of Section 3.1; to the left the version of mesh creation followed by sampling on fixed meshes, in Section 2.1, and to the right the path dependent sampling version in Section 2.2. The computational cost is measured as the total number of Euler time steps taken in all refinement iterations on all levels for all realizations. The graphs show three independent realizations of the underlying Wiener processes for each prescribed tolerance. A least squares fit, in $\log_2 - \log_2$-scale, of the model $cost = c_1 \mathrm{TOL}^{-c_2}(1 + \log_2(\mathrm{TOL}_{\mathrm{T},0}/\mathrm{TOL}))^2$ gives $c_2 = 1.7$ and $c_2 = 1.8$ in the two cases respectively; this is slightly better than the prediction of Theorem 3 of Section 4

The pronounced clustering of data points in the left graph is primarily due to the fact that the tolerance changes by a constant factor $2^{1/4}$ and for this example the adaptive algorithm generates uniform meshes by repeated halving of the same initial uniform mesh; the effect is that several consecutive tolerances result in identical mesh hierarchies, while only the number of samples changes. With the pathwise adaptive algorithm, to the right, the generated meshes are again uniform for geometric Brownian motion, but the resolution may be different for each individual outcome of the Wiener process; again the number of levels will be constant for four consecutive tolerances since it is decided by the requirement that $\mathrm{TOL}_{\mathrm{T},0} = 2^L \mathrm{TOL}_{\mathrm{T},L}$ is smaller than a fixed constant.

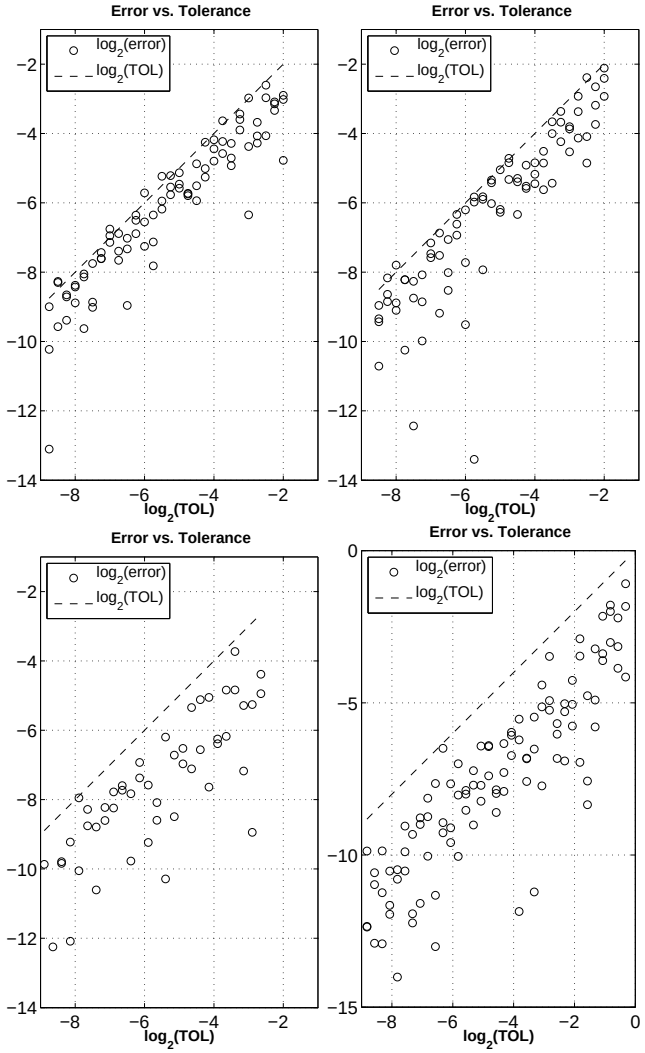


Figure 2: These accuracy tests show the error versus the prescribed tolerance when the adaptive MLMC algorithm is applied to the test examples of Section 3; to the left the version of Section 2.1 applied to the geometric Brownian motion in Section 3.1 (top) and the singularity problem in Section 3.2 (bottom), and to the right the version of Section 2.2 applied to the geometric Brownian motion in Section 3.1 (top) and the stopped diffusion problem in Section 3.3 (bottom).

| | | sampled randn's | | accepted Euler steps | | all Euler steps | |
|---|---|---|---|---|---|---|---|
| Problem | Version | $c_1$ | $c_2$ | $c_1$ | $c_2$ | $c_1$ | $c_2$ |
| GBM, | Sec. 2.1 | 6.9 | 1.7 | – | – | 7.0 | 1.7 |
| GBM, | Sec. 2.2 | 6.1 | 1.7 | 6.5 | 1.8 | 7.0 | 1.8 |
| Sing., | Sec. 2.1 | 13.2 | 2.0 | – | – | 13.4 | 2.0 |
| Barrier, | Sec. 2.2 | 8.2 | 1.9 | 8.8 | 1.9 | 9.5 | 2.1 |

Table 1: Complexity estimates for the three different problems: the geometric Brownian motion of Section 3.1, the deterministic singularity problem of Section 3.2, and the stopped diffusion problem of Section 3.3. The tabulated values are least square fits of the parameters $c_1$ and $c_2$ in the model
$\log_2(\mathit{cost}) = c_1 + \log_2\left(\mathrm{TOL}^{-c_2}(1+\log_2(\mathrm{TOL}_{\mathrm{T},0}/\mathrm{TOL}))^2\right)$
when the cost is measured in three different ways: by counting the total number of sampled random variables, by the number of accepted Euler steps
$\mathcal{A}[\mathit{cost}] = M_0\,\mathcal{A}(N_0;M_0) + \sum_{\ell=1}^{L} M_\ell\{\mathcal{A}(N_\ell;M_\ell)+\mathcal{A}(N_{\ell-1};M_\ell)\}$
which is approximated by the work estimate defined in (4.2), and by counting the total number of Euler steps performed when solving the primal problem in *all* refinement stages for all levels in the multilevel algorithms. The cost when counting Euler steps in the accepted meshes only was not recorded in the experiments using the path independent version of the algorithm, but it is by necessity bounded from below and above by the other two measures of the work.

## 3.2 Drift singularity, linear SDE

*Consider for a real constant $\alpha \in (0,T)$ the linear stochastic differential equation*

$$\begin{aligned} dX(t) &= \begin{cases} X(t)\,dW(t), & t\in[0,\alpha], \\ \frac{X(t)}{2\sqrt{t-\alpha}}\,dt + X(t)\,dW(t), & t\in(\alpha,T], \end{cases} \\ X(0) &= 1, \end{aligned} \tag{3.1}$$

*with the unique solution*

$$X(t) = \begin{cases} \exp\big(W(t)-t/2\big), & t\in[0,\alpha], \\ \exp\big(W(t)-t/2\big)\,\exp(\sqrt{t-\alpha}), & t\in(\alpha,T]. \end{cases}$$

*The goal is to approximate the expected value* $\mathrm{E}[X(T)] = \exp(\sqrt{T-\alpha})$. Here we choose $T=1$ and $\alpha = T/3$. To avoid evaluating arbitrarily large values of the drift in (3.1) we modify the drift to be

$$a(t,x) = \begin{cases} 0, & t\in[0,\alpha], \\ \frac{x}{2\sqrt{t-\alpha+\mathrm{TOL}^4}}, & t\in(\alpha,T], \end{cases} \tag{3.2}$$

yielding a higher order perturbation $\mathcal{O}\left(\mathrm{TOL}^2\right)$ in the computed result and in the size of the optimal time steps. This regularization was applied to maintain consistency with the numerical tests in [19], but is not strictly necessary with

the upper bound, $\rho \leq \rho_{up}(\mathrm{TOL})$, on the error density in (1.24). Due to the time discontinuity of the drift function and to ensure optimal convergence of the adaptive algorithms, we modify the Euler method to

$$\overline{X}_{n+1} - \overline{X}_n = a(\hat{t}, \overline{X}_n)\,\Delta t_n + \overline{X}_n\,\Delta W_n, \qquad n = 0, 1, 2, \ldots, \tag{3.3}$$

where we choose the stochastic evaluation time $\hat{t} \in \{t_n, t_{n+1}\}$ so that

$$\left|a(\hat{t}, \overline{X}_n)\right| = \max\left(\left|a(t_n, \overline{X}_n)\right|, \left|a(t_{n+1}, \overline{X}_n)\right|\right).$$

Observe that the use of $\hat{t}$ does not change the adapted nature of the Euler method.

Since the added difficulty compared to example 3.1 is a singularity in the drift at a deterministic time, the path independent adaptive algorithm described in Section 2.1 is the most suitable, and it is used in this example. The goal here is to verify that the adaptive multilevel algorithms of Section 2 give the same improvement from the single level adaptive algorithm as multilevel Monte Carlo does in the uniform case for regular problems.

The accuracy test in Figure 2 shows good agreement between observed error and prescribed tolerance. As shown in the complexity study in Table 1 and Figure 3 the computational costs grow like $\mathrm{TOL}^{-1.8}(1+\log\left(\mathrm{TOL}_{\mathrm{T},0}/\mathrm{TOL}\right))^2$ which is slightly better than the predicted complexity $\mathrm{TOL}^{-2}(\log\left(\mathrm{TOL}_{\mathrm{T},0}/\mathrm{TOL}\right))^2$. The cost of the mesh construction phase of the algorithm is seen to be negligible compared to the total work.

In this example the weak rate of convergence for the Euler–Maruyama method with uniform time steps is only 1/2, so the total cost for a single level uniform time stepping algorithm is proportional to $\mathrm{TOL}^{-4}$. The left part of Figure 4 shows that the single level version of the adaptive algorithm improves that complexity to approximately $\mathrm{TOL}^{-3}$, while the multilevel version improves the complexity by nearly one order more. With the regularization (3.2) the observed order of strong convergence of the Euler–Maruyama method with uniform time steps is still 1/2, so the complexity estimate in Theorem 1 of [7] for uniform multilevel simulations applies, and we should get the ideal complexity $(\epsilon^{-1}\log\left(\epsilon^{-1}\right))^2$ for a mean square error of size $\epsilon^2$. The right part of Figure 4 shows that this is approximately true for the cost as a function of the maximal observed error over 11 independent realizations.

## 3.3 Stopped diffusion

Here we compute the solution of a more challenging problem that motivates the use of stochastic time steps that are adaptively refined for each sample path.

The additional difficulty of the problem is that we now wish to compute approximations of an expected value

$$\mathrm{E}[g(X(\tau), \tau)], \tag{3.4}$$

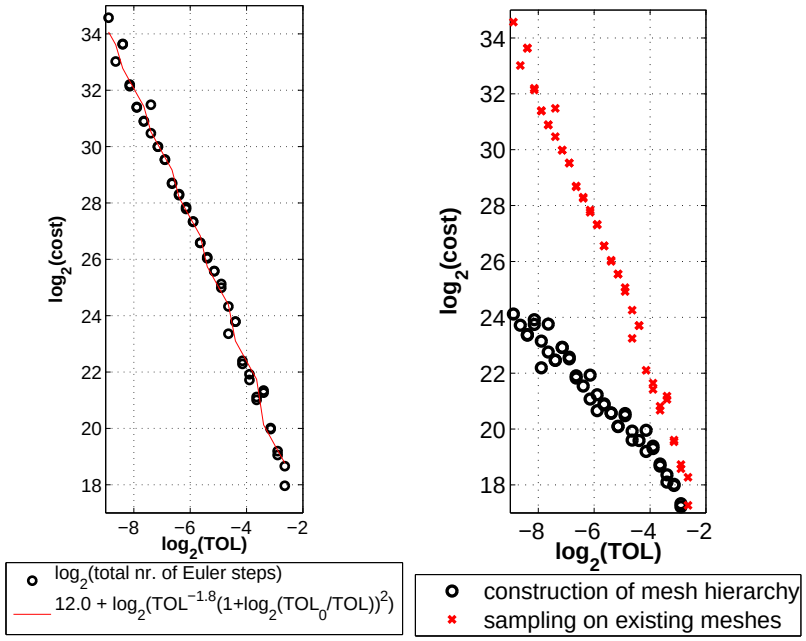


Figure 3: Experimental complexity when the algorithm in Section 2.1 is applied to the drift singularity problem in Section 3.2. To the left is shown the cost of both phases of the algorithm, and to the right the contribution from the generation of the mesh hierarchy and the subsequent sampling to reduce the statistical error; it is clear that the cost of the first phase is negligible compared to the second for small tolerances. The computational cost is measured as the total number of Euler time steps taken in all refinement iterations on all levels for all realizations. The graphs show three independent realizations of the underlying Wiener processes for each prescribed tolerance. A least squares fit, in $\log_2 - \log_2$-scale, of the model $cost = c_1 \mathrm{TOL}^{-c_2}(1 + \log_2(\mathrm{TOL}_{\mathrm{T},0}/\mathrm{TOL}))^2$ gives $c_2 = 1.8$.

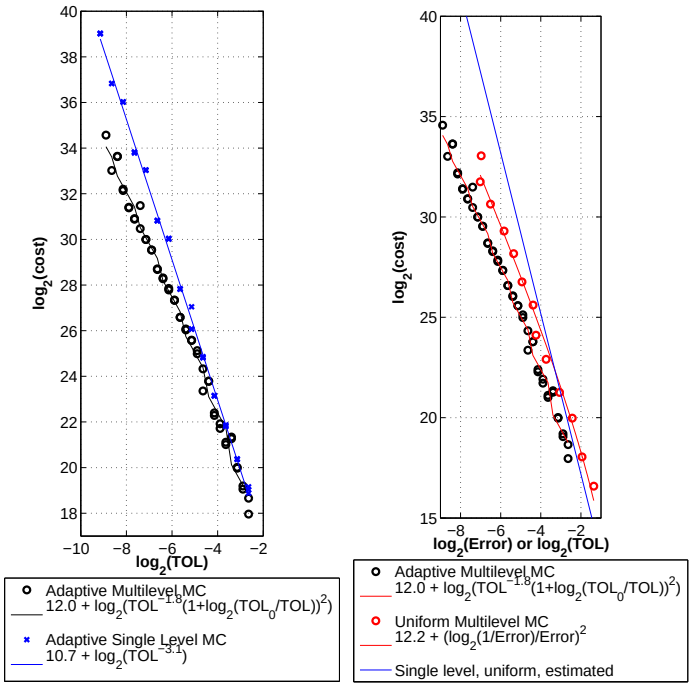


Figure 4: The computational cost of the path independent adaptive algorithm of Section 2.1, applied to the deterministic singularity problem 3.2, is compared to several alternatives. Left: the multilevel version improves the computational complexity of the single level version of the same adaptive algorithm from approximately proportional to $\mathrm{TOL}^{-3}$ to approximately proportional to $\mathrm{TOL}^{-2}\log\left(\mathrm{TOL}^{-1}\right)^2$. Right: the cost of a standard, uniform time step, Monte Carlo method would be proportional to $\mathrm{TOL}^{-4}$; here the work was estimated from a Central Limit Theorem type confidence interval based on the time discretization errors and sample variances. The cost of the uniform MLMC method is shown as a function of the maximal error over 11 realizations. The observed cost oscillates around $(\epsilon^{-1}\log\left(\epsilon^{-1}\right))^2$, which is expected since the observed observed strong order of convergence is $1/2$.
For the adaptive algorithm the cost is estimated by the total number of Euler steps taken on all levels in all stages of the adaptive refinement process.

where $X(t)$ solves the SDE (1.1) as before, but where the function $g : D \times [0,T] \to \mathbb{R}$ is evaluated at the first exit time

$$\tau := \inf\{t > 0 : (X(t), t) \notin D \times (0,T)\}$$

from a given open domain $D \times (0,T) \subset \mathbb{R}^d \times (0,T)$. This kind of stopped (or killed) diffusion problems arise for example in mathematical finance when pricing barrier options and for boundary value problems in physics.

The main difficulty in the approximation of the stopped diffusion on the boundary $\partial D$ is that a continuous sample path may exit the given domain $D$ even though a discrete approximate solution does not cross the boundary of $D$. Due to this hitting of the boundary the order of weak convergence of the Euler–Maruyama method is reduced from 1 to 1/2, in terms of the step size of uniform meshes; see [10]. In this subsection we combine the adaptive multilevel algorithm of Section 2.2 with an error estimate derived in [5] that also takes into account the hitting error. This error estimate, and the adaptive algorithm, can be used also when $D$ is multi dimensional even if the boundary $\partial D$ has corners for example.

The hitting error is accounted for by an extra contribution to the error density in (1.23); this contribution can be expressed in terms of exit probabilities for individual time steps, conditioned on the computed path at the beginning and the end of the time steps, and of the change in the goal function, $g$, when evaluated at a possible exit point within the time step instead of the actually computed exit $(\overline{X}(\bar{\tau}), \bar{\tau})$. The full expression of the resulting error indicators is given in equation (50) of [5]. Since the differential $\partial_i g(\overline{X}(T), T)$ in the discrete dual backward problem (1.17) does not exist if $T$ is replaced by $\bar{\tau} < T$ this initial value must be alternatively defined; this can be done using difference quotients with restarted computed trajectories as described, both for the discrete dual and for its first and second variations, in equations (20-25) of [5]. Note that for this modified error density the proof in [20] of almost sure convergence to a limit density does not apply.

In addition to the modification of the error density a lower bound is introduced on the step size to avoid excessive refinements near the barrier,

$$\Delta t_n \geq \min\left\{ \mathrm{TOL}_{\mathrm{T},\ell}{}^{1.5}, \frac{\mathrm{dist}_n \mathrm{dist}_{n+1} / b(\overline{X}(t_n;\omega), t_n)^2}{-3\log\left(\mathrm{TOL}_{\mathrm{T},\ell}\right)} \right\}, \tag{3.5}$$

where $\mathrm{dist}_j$ denotes the distance from $\overline{X}(t_j;\omega)$ to the barrier.

*For the numerical example we consider the stopped diffusion problem*

$$\begin{aligned} dX(t) &= \frac{11}{36} X(t)\, dt + \frac{1}{6} X(t)\, dW(t), \quad \textit{for } t \in [0,2] \textit{ and } X(t) \in (-\infty, 2), \\ X(0) &= 1.6. \end{aligned} \tag{3.6}$$

*For $g(x,t) = x^3 e^{-t}$ with $x \in \mathbb{R}$, this problem has the exact solution $\mathrm{E}[g(X_\tau, \tau)] = u(X(0),0) = X(0)^3$, where the solution, $u$, of the Kolmogorov backward equation is $u(x,t) = x^3 e^{-t}$.* We chose an example in one space dimension for simplicity, although it is only in high dimension that Monte Carlo methods are more

efficient than deterministic finite difference or finite element methods to solve stopped diffusion problems. The comparison here between the standard Monte Carlo and the Multilevel Monte Carlo methods in the simple one dimensional example indicates that the Multilevel Monte Carlo method will also be more efficient in high dimensional stopped diffusion problems, where a Monte Carlo method is a good choice. In the case of a scalar SDE, where $D$ is an interval on the real line, the strong order of convergence of the Euler–Maruyama scheme for barrier problems can be close to 1/2. In fact, it is shown in [8]that $\mathrm{Var}\big(g(\overline{X}_\ell) - g(\overline{X}_{\ell-1})\big) = \mathcal{O}(\Delta t^{1-\delta})$, for any $\delta > 0$, using the Euler–Maruyama method with uniform step size $\Delta t$ on a class of options including some barrier options. In this case Theorem 3.1 of [7] tells us that, for any choice of $\delta > 0$, uniform MLMC simulations can be performed at a cost $\mathcal{O}(\epsilon^{-2(1+\delta)})$ for a mean square error of order $\epsilon^2$.

In the remainder of this section we present results on the accuracy and cost of the adaptive multilevel algorithm of Section 2.2, applied to (3.6), with the error estimate modified for the barrier problem, and with the lower bound (3.5) on the step size. The algorithm was applied with a sequence of tolerances with three simulations for each tolerance using different initial states in the pseudo random number generator. The observed errors are scattered below the corresponding tolerances in Figure 2, showing that the algorithm achieves the prescribed accuracy.

The experimental complexity is illustrated in Figure 5 and Table 1. A least squares fit of the model

$$cost = c_1 \left(\frac{1}{\mathrm{TOL}}\right)^{c_2} \left(1 + \log\left(\frac{\mathrm{TOL}_{\mathrm{T},0}}{\mathrm{TOL}}\right)\right)^2 \tag{3.7}$$

in the $\log_2$-$\log_2$-scale of the graph using equal weights on all data points gives $c_2 = 1.9$ when the work is estimated by the total number of time steps on all the *accepted meshes*; this is the measure of work that is estimated by (4.2) in Section 4. When all Euler steps, in all refinement stages are included, the least squares fit gives $c_2 = 2.1$, which is also close to the rate predicted in Theorem 3. However, the corresponding cost using the single level adaptive algorithm with just one data point per tolerance used grows faster than $\mathrm{TOL}^{-3}$ in this example; see Figure 6.

In conclusion the observed convergence of the adaptive MLMC method applied to the barrier problem (3.6) is close to the predicted rate in Theorem 3. This shows an improved convergence compared to the single level version of the adaptive Monte Carlo algorithm where the cost grows approximately like $\mathrm{TOL}^{-3}$, which in itself is a better order of weak convergence than the one obtained using a single level Monte Carlo method with constant time steps where the cost grows like $\mathrm{TOL}^{-4}$.

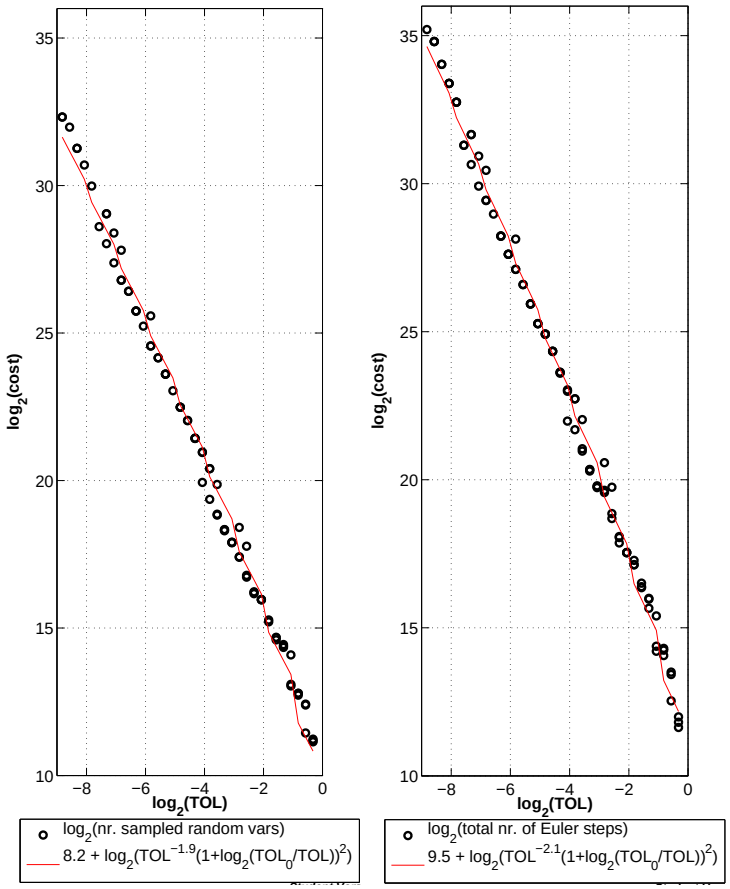


Figure 5: Experimental complexity for the barrier example in Section 3.3. The computational cost of the multilevel adaptive algorithm is shown for varying tolerances using three different initial states in the pseudo random number algorithm. To the left is shown the work estimate based on the number of Euler steps in the accepted meshes, $\mathcal{A}[cost] = M_0\,\mathcal{A}(N_0; M_0) + \sum_{\ell=1}^{L} M_\ell\{\mathcal{A}(N_\ell; M_\ell) + \mathcal{A}(N_{\ell-1}; M_\ell)\}$, which is the work measure closest to (4.2) used in Section 4; to the right is shown the estimate based on all Euler steps taken in all stages in the adaptive mesh refinement process. A least squares fit, in $\log_2$-$\log_2$-scale, of the model $cost = c_1 \mathrm{TOL}^{-c_2}(\log_2(1 + \mathrm{TOL}_{\mathrm{T},0}/\mathrm{TOL}))^2$ with equal weight on all observations results in $c_2 = 1.9$ and $c_2 = 2.1$ in the two cases. This is in agreement with the prediction in Theorem 3.

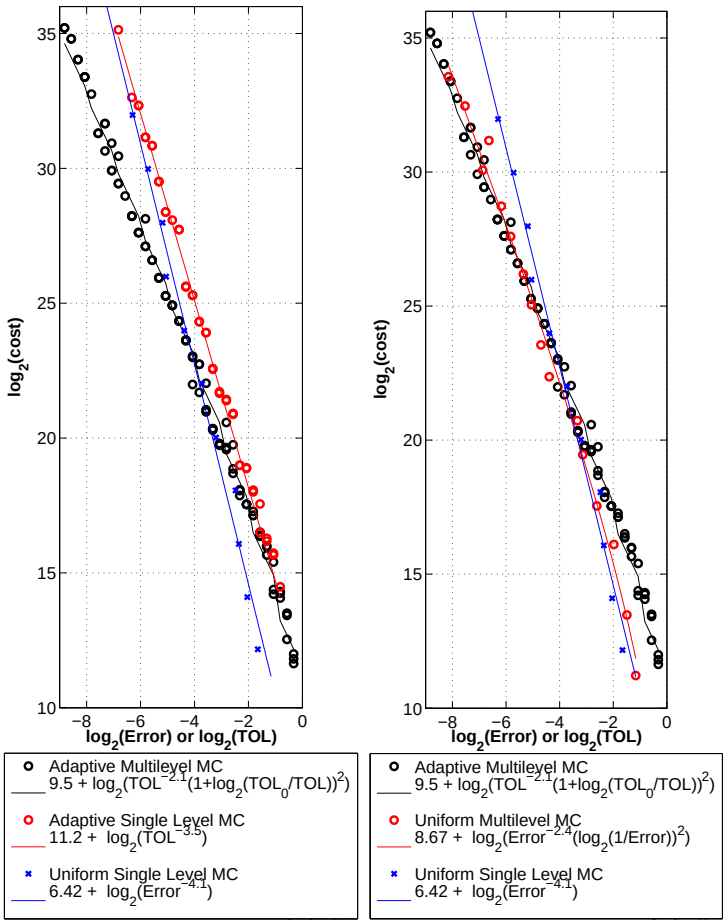


Figure 6: Left: The multilevel version of the path dependent adaptive algorithm of Section 2.2 applied to the barrier problem 3.3 improves the computational complexity of the single level version of the same adaptive algorithm; a single level method based on uniform time steps has even worse complexity with the computational cost growing like $\mathrm{TOL}^{-4}$. Right: The cost of the uniform MLMC method is shown as a function of the maximal error over 16 realizations. The observed cost is close to that of adaptive multilevel Monte Carlo, which is expected since the observed observed strong order of convergence is 1/2, but oscillates around a slightly worse fitted complexity $\epsilon^{-2.4}(\log(\epsilon^{-1}))^2$.
The cost is estimated by the total number of Euler steps taken on all levels in all stages of the adaptive refinement process.

# 4 Theoretical results

In this section we will study the asymptotic accuracy and complexity of the MLMC algorithm in the setting of stochastic adaptive time steps introduced in Section 2.2. We recall that for a sought accuracy $\mathrm{TOL} > 0$, the goal of the MLMC algorithm is to construct a Monte Carlo approximation of $\mathrm{E}[g(X(T))]$ that with probability close to one fulfills

$$\left|\mathrm{E}[g(X(T))] - \mathcal{A}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big); M_0\big)\right| \leq \mathrm{TOL}.$$

Our main result on asymptotic accuracy for MLMC algorithm, proved in Subsection 4.2, is

**Theorem 2** (Multilevel accuracy)**.** *Suppose that the modeling assumptions of Lemma 1 hold, that* (4.7) *holds, and that* $\mathrm{TOL_T} \leq \mathrm{TOL_S}$*. Then the adaptive MLMC algorithm with confidence parameter* $C_C > 0$ *and stochastic time steps* (2.20) *and* (2.21) *satisfies*

$$\liminf_{\mathrm{TOL}\downarrow 0} P\left(\left|\mathrm{E}[g(X(T))] - \mathcal{A}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big); M_0\big)\right| \leq \mathrm{TOL}\right) \geq \int_{-C_C}^{C_C} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx. \tag{4.1}$$

The motivation for introducing multiple levels in the MC algorithm is to reduce the computational complexity. To analyze the asymptotic complexity of the adaptive MLMC algorithm we define

$$\mathrm{WORK}(\mathrm{TOL}) = \sum_{\ell=0}^{L} \mathrm{E}[M_\ell]\mathrm{E}[N_\ell], \tag{4.2}$$

recalling that $M_\ell$ denotes the number of realization samples $g(\overline{X}_\ell(T;\omega))$ at level $\ell$ required to control the statistical error, and $N_\ell$ denotes the number of adaptive time steps required in the construction of a numerical realization $g(\overline{X}_\ell(T;\omega))$ to control the time discretization error at level $\ell$. The function WORK(TOL) is thus our estimate of the average number of arithmetic operations required in the generation and sampling of $\{g(\overline{X}_\ell(T))\}_{\ell=0}^{L}$ to approximate $\mathrm{E}[g(X(T))]$ for the prescribed confidence $C_C$ and accuracy TOL. By analyzing the asymptotics of $\mathrm{E}[M_\ell]$ and $\mathrm{E}[N_\ell]$ separately, our main complexity theorem is as follows.

**Theorem 3** (Multilevel computational complexity)**.** *Suppose the assumptions of Lemma 1 and* (4.7) *hold and suppose the lower bound for the error density is on the form* $\rho_{low}(\mathrm{TOL_T}) = \mathrm{TOL_T^{\bar{\gamma}}}$*, cf.* (1.24)*. Then the work for the MLMC algorithm using stochastic time steps, as defined in* (4.2)*, fulfills the following bounds:*

*(I) If* $\rho_{low}(\mathrm{TOL_T}) = \rho_{min} \in \mathbb{R}_+$ *(i.e.* $\bar{\gamma} = 0$*) and*

$$\min_{\tau\in[0,T]} |\widehat{\rho}(\tau)| \geq \rho_{min} \text{ a.s.} \tag{4.3}$$

*then*

$$\limsup_{\mathrm{TOL}\downarrow 0} \frac{\mathrm{WORK(TOL)TOL}^2}{L^2} \leq \frac{16\, C_C^2\, C_G}{\mathrm{TOL}_{\mathrm{T},Max}\, C_R} \left( \mathrm{E}\left[ \int_0^T \sqrt{|\widehat{\rho}(\tau)|} d\tau \right] \right)^2 . \tag{4.4}$$

*(II) Let* $\bar{\gamma}$ *be constant valued and satisfying* $0 < \bar{\gamma} < \alpha/(2+\alpha)$ *with* $0 < \alpha < 1/2$, *cf.* (1.24). *Then*

$$\limsup_{\mathrm{TOL}\downarrow 0} \frac{\mathrm{WORK(TOL)TOL}^{2+\bar{\gamma}}}{L} \leq \frac{(2+\bar{\gamma})^{2+\bar{\gamma}} 2^{2+\bar{\gamma}}}{\bar{\gamma}^{1+\bar{\gamma}} \log(2)} \cdot \frac{C_C^2\, C_G\, C_S^{\bar{\gamma}}}{\left(\mathrm{TOL}_{\mathrm{T},Max}\right)^{1-\bar{\gamma}}\, C_R} \left( \mathrm{E}\left[ \int_0^T \sqrt{|\widehat{\rho}(\tau)|} d\tau \right] \right)^2 . \tag{4.5}$$

*(III) If* $\bar{\gamma} \to 0$ *and* $L\bar{\gamma} \to \infty$ *as* $\mathrm{TOL} \downarrow 0$, *then*

$$\limsup_{\mathrm{TOL}\downarrow 0} \frac{\mathrm{WORK(TOL)TOL}^2\, \bar{\gamma}}{L\, 2^{\bar{\gamma}L}} \leq \frac{16\, C_C^2\, C_G}{\log(2)\, \mathrm{TOL}_{\mathrm{T},Max}\, C_R} \left( \mathrm{E}\left[ \int_0^T \sqrt{|\widehat{\rho}(\tau)|} d\tau \right] \right)^2 . \tag{4.6}$$

*Here, the number of levels* $L = \mathcal{O}\left(\log(\mathrm{TOL}^{-1})\right)$, $C_C$ *is the confidence parameter,* $C_R$ *and* $C_S$ *are refinement parameters described by* (2.20) *and* (2.21), $C_G$ *is the constant in the second moment bound* (4.43), *where* $\mathrm{TOL}_{\mathrm{T},Max}$ *is the upper bound of the time discretization tolerance at level* $\ell = 0$, *and* $\bar{\gamma}$ *is the lower bound error density exponent;* $\rho_{low}(\mathrm{TOL_T}) = \mathrm{TOL}_{\mathrm{T}}^{\bar{\gamma}}$, *cf.* (1.24).

**Remark 1** (Complexity example)**.** *Case (III) of Theorem 3 implies that if the exponent of the lower error density* $\rho_{low}$ *is given by* $\bar{\gamma}(\mathrm{TOL}) = \log_2(\log_2(L))/L$, *then the following complexity bound, notably close to the standard complexity in the setting of uniform time steps, is achieved:*

$$\limsup_{\mathrm{TOL}\downarrow 0} \frac{\mathrm{WORK(TOL)\, TOL}^2\, \log_2(\log_2(L))}{L^2\, \log_2(L)} \leq \frac{16\, C_C^2\, C_G}{\log(2)\, \mathrm{TOL}_{\mathrm{T},Max}\, C_R} \left( \mathrm{E}\left[ \int_0^T \sqrt{|\widehat{\rho}(\tau)|} d\tau \right] \right)^2 .$$

To introduce the reader gently to our proofs of Theorem 2 and 3, we have chosen to first prove analogous results for the SLMC algorithm in Subsection 4.1. With the single level proofs fresh in mind, we move on to the more daunting task of proving Theorem 2 and 3 in Subsection 4.2. We restrict ourselves to proving Theorem 2 and 3 for the stochastic time step setting only. The stochastic time step setting is however the most general setting, so one can easily prove corresponding results for the deterministic time step setting as well.

In addition to Lemma 1, the analysis in this section will be derived relying on one more technical assumption on strong approximations.

For $p = 2$ and 4, we have that

$$\mathrm{E}\Big[\big|g(X(T)) - g\big(\overline{X}(T)\big)\big|^p\Big] = \mathcal{O}\left(\frac{\mathrm{TOL}_\mathrm{T}}{\rho_{low}(\mathrm{TOL}_\mathrm{T})}\right)^{p/2} \qquad (4.7)$$
$$\mathrm{E}\Big[\big|g\big(\overline{X}(T)\big)\big|^p\Big] = \mathcal{O}\left(1\right),$$

with $\rho_{low}(\mathrm{TOL}_\mathrm{T})$ denoting the lower bound for the error density, as defined in (1.24). The work [24] gives conditions under which (4.7) is fulfilled.

## 4.1 Single level results

The SLMC algorithm we consider in this subsection, first described and analyzed in [24], is reanalyzed here with the goal to construct proofs for the asymptotic accuracy and complexity of the SLMC algorithm that are later extended to the MLMC algorithm. In the first lemma we show that the adaptive refinement Algorithm 7 stops after a finite number of iterations. This property allows us to later bound the amount of computational work in the single level adaptive algorithm. It also has another important implication: the imposed lower bound on the error density, $\rho_{low}(\mathrm{TOL}_\mathrm{T})$ in (1.10), ensures that the maximum mesh size of the mesh generated by Algorithm 7, $\Delta t_{\sup}(\mathrm{TOL}_\mathrm{T})$ introduced in Lemma 2, tends to zero as $\mathrm{TOL}_\mathrm{T}$ tends to zero. This in turn implies the almost sure convergence of the error density, which is crucial in the proofs of the main results of this section. A similar result for the multilevel case is direct to obtain and will not be stated for the sake of brevity.

**Lemma 3** (Stopping)**.** *Suppose the adaptive Algorithm 7 applies the mesh refinement strategy* (2.20) *and* (2.21) *on a set of realizations having the same uniform initial mesh of step size $\Delta t_0$. Further, assume that the estimated average number of time steps, $\overline{\mathcal{N}}_{in}$, satisfies*

$$\overline{\mathcal{N}}_{in} < N_{up} := \frac{T^2 \rho_{up}(\mathrm{TOL}_\mathrm{T})}{C_R\,\mathrm{TOL}_\mathrm{T}}. \qquad (4.8)$$

*Then, given a prescribed accuracy parameter $\mathrm{TOL}_\mathrm{T} > 0$, the adaptive refinement Algorithm 7 stops after a finite number of iterations.*

*Proof.* The main idea of the proof is to use the uniform upper bound on the error density, $\bar{\rho} \leq \rho_{up}(\mathrm{TOL}_\mathrm{T})$, according to (1.10). Given an initial mesh size $\Delta t_0$ with $N_0$ time steps, Algorithm 7 satisfies both the stopping and the non-refinement conditions, (2.20) and (2.21), for the uniform mesh size

$$\tilde{\Delta} t(\mathrm{TOL}_\mathrm{T}) = \frac{\Delta t_0}{\max\{1, 2^k\}}, \text{ with } k = \left\lceil \log_2\left(\frac{\rho_{up}(\mathrm{TOL}_\mathrm{T})\,T\,\Delta t_0}{C_\mathrm{R}\,\mathrm{TOL}_\mathrm{T}}\right)\right\rceil. \qquad (4.9)$$

Furthermore, if during the refinement process one time step reaches the mesh size $\tilde{\Delta} t(\mathrm{TOL}_\mathrm{T})$ then it cannot be refined further, according to (2.21) and (4.9).

Since the number of possible refinements from the initial mesh size $\Delta t_0$ to the mesh $\tilde{\Delta}t(\mathrm{TOL_T})$ is bounded by $N_0 \max\{1, 2^k\}$, there is only a finite number of possible refinements. The proof is concluded by observing that the Algorithm 7 either stops or makes at least one refinement during each iteration. □

The work [20] also proves a similar stopping result, cf. Theorem 3.2 in [20], based on the assumption that the initial mesh is sufficiently refined so that the error density does not vary too much between refinement levels. Then, when the single level adaptive algorithm stops, one can prove asymptotic accuracy and efficiency estimates on the resulting approximation. In contrast, here we make essentially no assumption on the initial mesh size $\Delta t_0$: although the quality of the resulting approximation for the lower levels of the multilevel estimator may be poor, they have no influence in the bias of the multilevel approximation, which is only determined by the finest level, $L$. Since $L \to \infty$ as $\mathrm{TOL} \downarrow 0$ we can still prove asymptotic accuracy and efficiency estimates. Finally, we observe that assumption (4.8) is fulfilled in all practical cases since one should start the adaptive algorithm with $\overline{\mathcal{N}}_{in}$ of the order of $\mathrm{TOL_T}^{-1}$, which is much smaller than $N_{up}$.

The following proofs are inspired by the treatment by Chow and Robbins [3] on the accuracy and complexity of sequential stopping rules for sampling i.i.d. random variables.

We denote the SLMC sample average estimator of $\mathrm{E}[g(X(T))]$ by

$$\mathcal{A}\big(g(\overline{X}(T)); M\big) = \sum_{i=1}^{M} \frac{g\left(\overline{X}(T;\omega_i)\right)}{M},$$

where the realizations of $\overline{X}(T)$ are generated on adaptive meshes and fulfill the weak error bound $\left|\mathrm{E}\left[g\big(\overline{X}(T)\big) - g(X(T))\right]\right| \lesssim \mathrm{TOL_T}$. Here the total tolerance TOL is split into a time discretization error tolerance and a statistical error tolerance, $\mathrm{TOL} = C_S\mathrm{TOL_T} + \mathrm{TOL_S}$ (Remark 2 discusses the optimal splitting). Let $2^{\mathbb{N}}$ denote the set $\{2^n | n \in \mathbb{N}\}$. For the SLMC estimator, the number of samples used in the sample average estimator to control the statistical error $\left|\mathcal{A}\big(g(\overline{X}(T)); M\big) - \mathrm{E}\left[g\big(\overline{X}(T)\big)\right]\right| \leq \mathrm{TOL_S}$ is a stochastic process $M : \mathbb{R}_+ \to 2^{\mathbb{N}}$ defined by

$$\begin{aligned} M(\mathrm{TOL_S}) := &\text{ the smallest } k \in 2^{\mathbb{N}+\lceil \log_2(\mathrm{TOL}^{-1})\rceil} \\ &\text{such that } \mathcal{V}\big(g\big(\overline{X}(T)\big); k\big) < \frac{k{\mathrm{TOL_S}}^2}{C_C^2}, \end{aligned} \tag{4.10}$$

where the sample variance is defined by

$$\mathcal{V}\big(g\big(\overline{X}(T)\big); k\big) = \sum_{i=1}^{k} \frac{\big(g(\overline{X}(T;\omega_i)) - \mathcal{A}\big(g(\overline{X}(T)); k\big)\big)^2}{k-1}. \tag{4.11}$$

Restricting the initial value of $M$ to the set $2^{\mathbb{N}+\lceil \log_2(\mathrm{TOL}^{-1})\rceil}$ implies that that $\lim_{\mathrm{TOL}\downarrow 0} M = \infty$. The asymptotic behavior of $M$ as $\mathrm{TOL} \downarrow 0$ is crucial in our

proofs of the asymptotic accuracy and complexity. When proving the asymptotically accuracy result of Theorem 4, $M$ should increase sufficiently fast to obtain the sought confidence. For the complexity result of Theorem 5, it is on the other hand useful to bound $M$ from above and ensure that it does not grow too fast.

**Lemma 4.** *Suppose the assumptions of Lemma 1 hold and* (4.7) *holds for at least* $p = 2$. *Then*

$$\liminf_{\mathrm{TOL}\downarrow 0} \frac{M\mathrm{TOL}_S^2}{\mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)C_C^2} = 1 \quad \text{a.s.} \quad \textit{and} \quad \limsup_{\mathrm{TOL}\downarrow 0} \frac{M\mathrm{TOL}_S^2}{\mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)C_C^2} = 2 \quad \text{a.s.} \tag{4.12}$$

*Proof.* The strong convergence (4.7) for $p = 2$, gives $\lim_{\mathrm{TOL}\downarrow 0} Var(g\big(\overline{X}(T)\big)) = Var(g(X(T)))$, which in particular means that there exists a constant $\widetilde{\mathrm{TOL}} > 0$ such that

$$\frac{\mathrm{Var}(g(X(T)))}{2} < \mathrm{Var}\big(g\big(\overline{X}(T)\big)\big) < 2\mathrm{Var}(g(X(T))), \qquad \forall \mathrm{TOL} \in (0, \widetilde{\mathrm{TOL}}]. \tag{4.13}$$

The strong law of large numbers then implies that for all

$$\lim_{k\to\infty} \mathcal{V}\big(g\big(\overline{X}(T)\big); k\big) = \mathrm{Var}\big(g\big(\overline{X}(T)\big)\big) \quad \text{a.s.} \quad \forall \mathrm{TOL} \in (0, \widetilde{\mathrm{TOL}}]. \tag{4.14}$$

In order to prove (4.12), introduce the sequence of stochastic processes $y_k : \mathbb{R}_+ \to \mathbb{R}_+$ sub-indexed by $k \in 2^{\mathbb{N}+\lceil \log_2(\mathrm{TOL}^{-1})\rceil}$ and defined by

$$y_k(\mathrm{TOL}) = \frac{\mathcal{V}\big(g\big(\overline{X}(T)\big); k\big)}{\mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)}. \tag{4.15}$$

Using $y_k$, definition (4.10) of $M(\mathrm{TOL}_S)$ is equivalent to

$$\begin{aligned} M(\mathrm{TOL}_S) := &\text{ the smallest } k \in 2^{\mathbb{N}+\lceil \log_2(\mathrm{TOL}^{-1})\rceil} \\ &\text{such that } y_k(\mathrm{TOL}_S) < \frac{k\mathrm{TOL}_S^2}{\mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)C_C^2}. \end{aligned}$$

This gives rise to the bounds

$$y_M(\mathrm{TOL}_S) < \frac{M\mathrm{TOL}_S^2}{\mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)C_C^2} \le 2y_{M/2}(\mathrm{TOL}_S). \tag{4.16}$$

Combining (4.14) with definition (4.10), which ensures that $\lim_{\mathrm{TOL}\downarrow 0} M = \infty$, we conclude that

$$\lim_{\mathrm{TOL}\downarrow 0} \mathcal{V}\big(g\big(\overline{X}(T)\big); M(\mathrm{TOL}_S)\big) = \mathrm{Var}(g(X(T))) > 0 \quad \text{a.s.}$$

which implies that also $\lim_{\mathrm{TOL}\downarrow 0} y_M(\mathrm{TOL}_S) = 1$ a.s. Statement (4.12) then follows by taking limits in (4.16). □

Having obtained asymptotic bounds for $M$, we are ready to prove the main accuracy result for the SLMC algorithm.

**Theorem 4** (Single level accuracy)**.** *Suppose the modeling assumptions of Lemma 1 holds, that* (4.7) *holds for at least* $p = 2$*, and that* $\mathrm{TOL_T} \leq \mathrm{TOL_S}$*. Then, the adaptive SLMC algorithm with confidence refinement parameter* $C_C > 0$*, and stochastic time steps* (2.20) *and* (2.21)*, satisfies*

$$\liminf_{\mathrm{TOL}\downarrow 0} P\left(\left|E\left[g(X(T))\right] - \mathcal{A}\big(g(\overline{X}(T)); M\big)\right| \leq \mathrm{TOL}\right) \geq \int_{-C_C}^{C_C} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx. \quad (4.17)$$

*Proof.* For a given $\delta > 0$, we first bound the probability in (4.17) from below as follows:

$$\begin{aligned}
&\liminf_{\mathrm{TOL}\downarrow 0} P\left(\left|\mathrm{E}[g(X(T))] - \mathcal{A}\big(g(\overline{X}(T)); M\big)\right| \leq \mathrm{TOL}\right) \\
&\geq \liminf_{\mathrm{TOL}\downarrow 0} P\Big(\left|\mathrm{E}\left[g(X(T)) - g\big(\overline{X}(T)\big)\right]\right| \\
&\qquad + \left|\mathrm{E}\left[g\big(\overline{X}(T)\big)\right] - \mathcal{A}\big(g(\overline{X}(T)); M\big)\right| \leq C_S \mathrm{TOL}_T + \mathrm{TOL}_S\Big) \\
&\geq \liminf_{\mathrm{TOL}\downarrow 0} P\Big(\left|\mathrm{E}\left[g(X(T)) - g\big(\overline{X}(T)\big)\right]\right| \leq (C_S + \delta)\mathrm{TOL}_T \\
&\qquad\qquad \text{and } \left|\mathrm{E}\left[g\big(\overline{X}(T)\big)\right] - \mathcal{A}\big(g(\overline{X}(T)); M\big)\right| \leq (1-\delta)\mathrm{TOL}_S\Big) \\
&= \liminf_{\mathrm{TOL}\downarrow 0} P\Big(\left|\mathrm{E}\left[g(X(T)) - g\big(\overline{X}(T)\big)\right]\right| \leq (C_S + \delta)\mathrm{TOL}_T\Big) \\
&\qquad\qquad \times P\left(\left|\mathrm{E}\left[g\big(\overline{X}(T)\big)\right] - \mathcal{A}\big(g(\overline{X}(T)); M\big)\right| \leq (1-\delta)\mathrm{TOL}_S\right)
\end{aligned} \quad (4.18)$$

The proof is continued by analyzing the two product terms of the last line of the inequality above separately:

**The time discretization error.** From the treatment of the time discretization error for the single level case we refer to the proof of Theorem 3.4, p. 530 in [19] which shows that

$$\limsup_{\mathrm{TOL}\downarrow 0} \frac{\left|\mathrm{E}\left[g(X(T)) - g\big(\overline{X}(T)\big)\right]\right|}{\mathrm{TOL}_T} \leq C_S.$$

Thereby,

$$\liminf_{\mathrm{TOL}\downarrow 0} P\Big(\left|\mathrm{E}\left[g(X(T)) - g\big(\overline{X}(T)\big)\right]\right| \leq (C_S + \delta)\mathrm{TOL}_T\Big) = 1.$$

**The statistical error.** For the above introduced $\delta > 0$, define the family of sets

$$\Omega_\delta(\mathrm{TOL}_S) = \left\{ k \in 2^{\mathbb{N} + \lceil \log_2(\mathrm{TOL}^{-1})\rceil} \Big| 1 - \delta < \frac{k\mathrm{TOL}_S^2}{\mathrm{Var}\big(g\big(\overline{X}(T)\big)\big) C_C^2} \leq 2 + \delta \right\}. \quad (4.19)$$

By the convergence (4.12), we conclude that

$$\lim_{\mathrm{TOL}\downarrow 0} P(M \in \Omega_\delta) = 1.$$

Recall that for the SLMC algorithm, the number of samples $M$ is determined in the step prior to generating the output $\mathcal{A}\big(g(\overline{X}(T)); M\big)$, so that $M$ is independent from $\mathcal{A}\big(g(\overline{X}(T)); M\big)$. Using this independence property, Fatou's Lemma, and Lindeberg-Feller's version of the Central Limit Theorem, cf. Theorem 6, yield that

$$
\begin{aligned}
&\liminf_{\mathrm{TOL}\downarrow 0} P\left(\left|\mathrm{E}\left[g\big(\overline{X}(T)\big)\right] - \mathcal{A}\big(g(\overline{X}(T)); M\big)\right| \leq (1-\delta)\mathrm{TOL}_S\right) \\
&= \liminf_{\mathrm{TOL}\downarrow 0} \sum_{k\in 2^{\mathbb{N}+\lceil \log_2(\mathrm{TOL}^{-1})\rceil}} P\left(\left|\mathrm{E}\left[g\big(\overline{X}(T)\big)\right] - \mathcal{A}\big(g(\overline{X}(T)); k\big)\right| \leq (1-\delta)\mathrm{TOL}_S\right) P(M=k) \\
&\geq \liminf_{\mathrm{TOL}\downarrow 0} \sum_{k\in\Omega_\delta} P\left(\left|\mathrm{E}\left[g\big(\overline{X}(T)\big)\right] - \mathcal{A}\big(g(\overline{X}(T)); k\big)\right| \leq (1-\delta)\mathrm{TOL}_S\right) P(M=k) \\
&\quad + \sum_{k\in 2^{\mathbb{N}+\lceil \log_2(\mathrm{TOL}^{-1})\rceil}\setminus\Omega_\delta} \liminf_{\mathrm{TOL}\downarrow 0} P\left(\left|\mathrm{E}\left[g\big(\overline{X}(T)\big)\right] - \mathcal{A}\big(g(\overline{X}(T)); k\big)\right| \leq (1-\delta)\mathrm{TOL}_S\right) P(M=k) \\
&\geq \liminf_{\mathrm{TOL}\downarrow 0} \sum_{k\in\Omega_\delta} P\left(\sqrt{k}\frac{\left|\mathrm{E}\left[g\big(\overline{X}(T)\big)\right] - \mathcal{A}\big(g(\overline{X}(T)); k\big)\right|}{\mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)} \leq (1-\delta)^{3/2} C_C\right) P(M=k) \\
&\geq \int_{-(1-\delta)^{3/2}C_C}^{(1-\delta)^{3/2}C_C} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx.
\end{aligned}
\tag{4.20}
$$

The proof is finished by noting that the argument leading to inequality (4.20) is valid for all $\delta > 0$. □

We conclude this subsection with a complexity analysis of the SLMC algorithm. Similar to the definition of the work for the MLMC algorithm given in (4.2), we define the SLMC work by

$$
\mathrm{WORK}(\mathrm{TOL}) = \mathrm{E}[M]\mathrm{E}[N], \tag{4.21}
$$

where we recall that $M$ denotes the number of samples of $g\big(\overline{X}(T)\big)$ required to control the statistical error and $N$ denotes the number of adaptive time steps required in the construction of a numerical realization $g(\overline{X}(T;\omega))$ to control the time discretization error $\left|\mathrm{E}\left[g\big(\overline{X}(T)\big) - g(X(T))\right]\right| \leq \mathrm{TOL}_\mathrm{T}$. We start by bounding $\mathrm{E}[M]$.

**Lemma 5.** *Suppose the assumptions of Lemma 1 and* (4.7) *hold. Then the number of samples used in the approximation of* $\mathrm{E}[g(X(T))]$ *is bounded by*

$$
\limsup_{\mathrm{TOL}\downarrow 0} \frac{\mathrm{E}[M]\mathrm{TOL}_\mathrm{S}^2}{\mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)C_C^2} \leq 2. \tag{4.22}
$$

*Proof.* For a given $\delta > 0$, define the deterministic function

$$
\widetilde{M}(\mathrm{TOL}_S) = \min\left\{ k \in 2^{\mathbb{N}+\lceil \log_2(\mathrm{TOL}^{-1})\rceil} \Big| \frac{k\mathrm{TOL}_S^2}{\mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)C_C^2} > 1+\delta \right\}.
$$

Assuming TOL is sufficiently small so that (4.13) holds, the relation (4.16), the fourth moment bound (4.7) and k-Statistics bounds on the variance of the sample variance, cf. [17], yield

$$
\begin{aligned}
P(M = 2\widetilde{M}) &\leq P\left(\frac{\mathcal{V}\Big(g\big(\overline{X}(T)\big);\widetilde{M}\Big)}{\mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)} > \widetilde{M}\frac{\mathrm{TOL}_S^2}{\mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)C_C^2}\right)\\
&\leq P\left(\frac{\mathcal{V}\Big(g\big(\overline{X}(T)\big);\widetilde{M}\Big)}{\mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)} > 1+\delta\right)\\
&\leq P\left(|\mathcal{V}\Big(g\big(\overline{X}(T)\big);\widetilde{M}\Big) - \mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)| > \delta\mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)\right)\\
&\leq 2\mathrm{E}\left[\frac{\left|\mathcal{V}\Big(g\big(\overline{X}(T)\big);\widetilde{M}\Big) - \mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)\right|^2}{\delta^2\mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)^2}\right]\\
&< \frac{C}{\delta^2\widetilde{M}},
\end{aligned}
$$

Furthermore, for $\ell = 1, 2, \ldots$ we get that

$$
\begin{aligned}
P(M = 2^{\ell+1}\widetilde{M}) &\leq P\left(\left|\mathcal{V}\Big(g\big(\overline{X}(T)\big);2^\ell\widetilde{M}\Big) - \mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)\right| > 2^{\ell-1}\mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)\right)\\
&\leq 2\mathrm{E}\left[\frac{\left|\mathcal{V}\Big(g\big(\overline{X}(T)\big);2^\ell\widetilde{M}\Big) - \mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)\right|^2}{2^{2(\ell-1)}\mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)^2}\right]\\
&< \frac{C}{2^{2\ell}\widetilde{M}}.
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
\frac{\mathrm{E}[M]\mathrm{TOL}_\mathrm{S}^2}{\mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)C_C^2} &\leq \frac{\left[P(M \leq \widetilde{M}) + \sum_{\ell=1}^\infty 2^\ell P(M = 2^\ell\widetilde{M})\right]\widetilde{M}\mathrm{TOL}_\mathrm{S}^2}{\mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)C_C^2}\\
&\leq 2(1+\delta)\left[P(M \leq \widetilde{M}) + P(M = 2\widetilde{M}) + \sum_{\ell=1}^\infty 2^{\ell+1}P(M = 2^{\ell+1}\widetilde{M})\right] \quad (4.23)\\
&\leq 2(1+\delta)\left[P(M \leq \widetilde{M}) + \frac{C}{\delta^2\widetilde{M}} + \frac{C}{\widetilde{M}}\sum_{\ell=1}^\infty 2^{-\ell}\right].
\end{aligned}
$$

By taking limits in the above inequality, we obtain

$$
\limsup_{\mathrm{TOL}\downarrow 0} \frac{\mathrm{E}[M]\mathrm{TOL}_S^2}{\mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)C_C^2} \leq 2(1+\delta),
$$

and noting that this result holds for any $\delta > 0$, the proof is finished. □

For an asymptotic bound on $\mathrm{E}[N]$, we recall Theorem 3.5 of [19]. The bound given in this theorem is derived by studying the asymptotic form of the error indicators obtained by the stopping condition (2.20). The theorem further shows that up to a multiplicative constant, the mesh refinement scheme (2.20)-(2.21) yields stochastic meshes which are optimal in mean sense. The theorem is here stated as a lemma.

**Lemma 6** (Single level asymptotic average number of time steps)**.** *Suppose that the assumptions of Lemma 1 hold. Then the final number of adaptive steps generated by the algorithm* (2.20) *and* (2.21) *satisfies asymptotically*

$$\limsup_{\mathrm{TOL}\downarrow 0} \mathrm{TOL}_T \ \mathrm{E}[N] \leq \frac{4}{C_R} \left( \mathrm{E}\left[ \int_0^T \sqrt{|\hat{\rho}(t)|} dt \right] \right)^2 . \tag{4.24}$$

The product of the asymptotic upper bounds for $\mathrm{E}[M]$ and $\mathrm{E}[N]$ and an optimization of the choice of $\mathrm{TOL}_T$ and $\mathrm{TOL}_S$ gives the following upper bound on the computational complexity for the SLMC algorithm.

**Theorem 5** (SLMC computational complexity)**.** *Suppose the assumptions of Lemma 1 and* (4.7) *hold. Then the work for the SLMC algorithm with stochastic time steps, defined in* (4.21)*, satisfies*

$$\limsup_{\mathrm{TOL}\downarrow 0} \mathrm{WORK}(\mathrm{TOL})\mathrm{TOL}^3 \leq \frac{2 \cdot 3^3 \mathrm{Var}(g(X(T)))C_C^2 C_S}{C_R} \left( \mathrm{E}\left[ \int_0^T \sqrt{|\hat{\rho}(t)|} dt \right] \right)^2 , \tag{4.25}$$

*where $C_C$ is the confidence parameter and $C_R$ and $C_S$ are refinement parameters described by* (2.20) *and* (2.21)*.*

*Proof.* Lemma 5 and 6 straightforwardly yield the upper bound

$$\limsup_{\mathrm{TOL}\downarrow 0} \mathrm{WORK}(\mathrm{TOL})\ \mathrm{TOL_S}^2 \mathrm{TOL_T} \leq \frac{2^3 \mathrm{Var}(g(X(T)))C_C^2}{C_R} \left( \mathrm{E}\left[ \int_0^T \sqrt{|\hat{\rho}(t)|} dt \right] \right)^2 .$$

So $\mathrm{WORK}(\mathrm{TOL}) = \mathcal{O}\left(\mathrm{TOL_S}^{-2}\mathrm{TOL_T}^{-1}\right)$. Minimizing $\mathrm{TOL_S}^{-2}\mathrm{TOL_T}^{-1}$ subject to the restriction $C_S\mathrm{TOL_T} + \mathrm{TOL_S} = \mathrm{TOL}$, yields

$$\mathrm{TOL_T} = \frac{\mathrm{TOL}}{3C_S} \text{ and } \mathrm{TOL_S} = \frac{2\mathrm{TOL}}{3}.$$

These values for $\mathrm{TOL_T}$ and $\mathrm{TOL_S}$ lead to the upper bound (4.25). □

**Remark 2.** *The optimal choices of* $\mathrm{TOL_T}$ *and* $\mathrm{TOL_S}$ *for minimizing* $\mathrm{WORK}(\mathrm{TOL})$ *are derived in the proof of Theorem 5 to be*

$$\mathrm{TOL_T} = \frac{\mathrm{TOL}}{3C_S} \text{ and } \mathrm{TOL_S} = \frac{2\mathrm{TOL}}{3}.$$

## 4.2 multilevel results

We recall from the description of the MLMC algorithm in Section 2.2 that given an accuracy $\mathrm{TOL} = C_S\mathrm{TOL}_\mathrm{T} + \mathrm{TOL}_\mathrm{S}$, the MLMC algorithm generates realizations $g\big(\overline{X}_\ell(T)\big)$ on adaptive meshes fulfilling the weak error bounds $\big|\mathrm{E}\big[g\big(\overline{X}_\ell(T)\big) - g(X(T))\big]\big| \lesssim \mathrm{TOL}_{\mathrm{T},\ell}$ on the levels $\ell = 0, 1, \ldots, L$. The time discretization tolerance levels are given by $\mathrm{TOL}_{\mathrm{T},\ell} = 2^\ell \mathrm{TOL}_\mathrm{T}$, and the number of levels is set by $L = \lfloor \log_2(\mathrm{TOL}_{\mathrm{T,Max}}/\mathrm{TOL}_\mathrm{T})\rfloor$, where $\mathrm{TOL}_{\mathrm{T,Max}}$ is a predetermined max time discretization tolerance value, cf. (2.3). The MLMC sample average estimator of $\mathrm{E}[g(X(T))]$ is denoted by

$$\mathcal{A}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big); M_0\big) = \sum_{i=1}^{M_0} \frac{g(\overline{X}_0(T;\omega_{0,i}))}{M_0} + \sum_{\ell=1}^{L}\sum_{i=1}^{M_\ell} \frac{\Delta_\ell g(\overline{X}(T;\omega_{\ell,i}))}{M_\ell},$$

where $M_0 \in 2^{L+\lceil C_{\mathcal{MLC}} L\rceil}2^\mathbb{N}$ denotes the number of samples on the coarsest level with the constant $C_{\mathcal{MLC}} \in (0,1)$, and the number of samples on higher levels is expressed in terms of $M_0$ by the ratio

$$M_\ell = \frac{M_0}{2^L}\left\lceil 2^L \frac{\rho_{low}(\mathrm{TOL}_{\mathrm{T},0})\mathrm{TOL}_{\mathrm{T},\ell}}{\rho_{low}(\mathrm{TOL}_{\mathrm{T},\ell})\mathrm{TOL}_{\mathrm{T},0}}\right\rceil = \frac{M_0}{2^L}\left\lceil 2^{L+(\bar\gamma-1)\ell}\right\rceil \quad \ell = 1, 2, \ldots, L. \tag{4.26}$$

The number of samples at the coarsest level is a stochastic process $M_0 : \mathbb{R}_+ \to 2^{\mathbb{N}+L+\lceil C_{\mathcal{MLC}} L\rceil}$ defined by

$$\begin{aligned} &M_0(\mathrm{TOL}_\mathrm{S}) = \text{ the smallest } k_0 \in 2^{\mathbb{N}+L+\lceil C_{\mathcal{MLC}} L\rceil} \text{ such that} \\ &\mathcal{V}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big); k_0\big) < \frac{k_0 {\mathrm{TOL}_\mathrm{S}}^2}{C_C^2}, \end{aligned} \tag{4.27}$$

where

$$\begin{aligned} \mathcal{V}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big); k_0\big) &= \sum_{i=1}^{k_0} \frac{\big(g(\overline{X}_0(T;\omega_{0,i})) - \mathcal{A}\big(g(\overline{X}_0(T;\omega_{0,\cdot})); k_0\big)\big)^2}{k_0 - 1} \\ &+ \sum_{\ell=1}^{L} \frac{k_0}{k_\ell}\sum_{i=1}^{k_\ell} \frac{\big(\Delta_\ell g(\overline{X}(T;\omega_{\ell,i})) - \mathcal{A}\big(\Delta_\ell g(\overline{X}(T;\omega_{\ell,\cdot})); k_\ell\big)\big)^2}{k_\ell - 1} \\ &= \mathcal{V}\big(g\big(\overline{X}_0(T;\omega_{0,\cdot})\big); k_0\big) + 2^L \sum_{\ell=1}^{L} \frac{\mathcal{V}\big(\Delta_\ell g\big(\overline{X}_0(T;\omega_{\ell,\cdot})\big); k_\ell\big)}{\big\lceil 2^{L+\ell(\bar\gamma-1)}\big\rceil} \end{aligned} \tag{4.28}$$

and, analogous to the definition of $M_\ell$,

$$k_\ell := \frac{k_0}{2^L}\left\lceil 2^{L+(\bar\gamma-1)\ell}\right\rceil, \quad \ell = 1, 2, \ldots, L. \tag{4.29}$$

**Remark 3.** *In the analysis of the SLMC Algorithm, the requirement* $M_0 \in 2^{\mathbb{N}+\lceil \log(1/\mathrm{TOL})\rceil}$ *ensured that the number of samples used in the MC estimate fulfilled* $\liminf_{\mathrm{TOL}\downarrow 0} M = \infty$. *For the MLMC Algorithm, we analogously ensure that* $\liminf_{\mathrm{TOL}\downarrow 0} M_L = \infty$ *by requiring that* $M_0 \in 2^{\mathbb{N}+L+\lceil C_{\mathcal{MLC}} L\rceil}$ *for any positive constant* $C_{\mathcal{MLC}}$.

The stochastic process $M_0$ is defined in a similar way as the stochastic process $M$ was defined for SLMC algorithm, cf. (4.10). For the SLMC algorithm, asymptotic accuracy and complexity results were easily obtained by applying the asymptotic bounds of $M$, cf. Lemma 4. Applying the same strategy for the MLMC algorithm, we will derive asymptotic bounds for $M_0$ and use these bounds to prove the accuracy and complexity results of Theorem 2 and 3.

**Lemma 7** (Asymptotic bounds for $M_0$)**.** *Let*

$$\mathrm{Var}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big)\big) := \mathrm{Var}\big(g\big(\overline{X}_0(T)\big)\big) + 2^L \sum_{\ell=1}^{L} \frac{\mathrm{Var}\big(\Delta_\ell g\big(\overline{X}(T)\big)\big)}{\lceil 2^{L+\ell(\bar{\gamma}-1)} \rceil} \tag{4.30}$$

*and suppose the assumptions of Lemma 1 and* (4.7) *hold. Further, assume that* $\mathrm{Var}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big)\big) > 0$ *for all sufficiently small* $\mathrm{TOL} > 0$*. Then* $M_0(\mathrm{TOL}_\mathrm{S})$ *defined according to* (4.27) *fulfills*

$$\begin{aligned} \liminf_{\mathrm{TOL}\downarrow 0} \frac{M_0 \mathrm{TOL}_\mathrm{S}^2}{\mathrm{Var}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big)\big) C_C^2} &= 1 \quad \textit{in probability, and} \\ \limsup_{\mathrm{TOL}\downarrow 0} \frac{M_0 \mathrm{TOL}_\mathrm{S}^2}{\mathrm{Var}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big)\big) C_C^2} &= 2 \quad \textit{in probability.} \end{aligned} \tag{4.31}$$

*Proof.* The definition of $M_0$ given in (4.27) implies that the following inequalities hold:

$$\frac{\mathcal{V}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big); M_0\big)}{\mathrm{Var}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big)\big)} \leq \frac{M_0 \mathrm{TOL}_\mathrm{S}^2}{\mathrm{Var}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big)\big) C_C^2} \leq 2 \frac{\mathcal{V}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big); M_0/2\big)}{\mathrm{Var}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big)\big)}.$$

So to conclude the proof, we will show that

$$\lim_{\mathrm{TOL}\downarrow 0} \frac{\mathcal{V}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big); M_0\big)}{\mathrm{Var}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big)\big)} = 1, \text{ in probability.} \tag{4.32}$$

Define the deterministic function $\tilde{k}_0(\mathrm{TOL}_\mathrm{T}) = 2^{L(\mathrm{TOL}_\mathrm{T}) + \lceil C_{\mathcal{MLC}} L(\mathrm{TOL}_\mathrm{T}) \rceil + 1}$ and let $\{\tilde{k}_\ell\}_{\ell=1}^L$ be the corresponding level functions defined according to (4.29).

Then, for a given $\epsilon > 0$, let us consider

$$
\begin{aligned}
&P\left(\left|\frac{\mathcal{V}_{\mathcal{MLC}}\Big(g\big(\overline{X}(T)\big);\tilde{k}_0\Big)}{\mathrm{Var}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big)\big)}-1\right|>\epsilon\right)\\
&=P\left(\left|\mathcal{V}_{\mathcal{MLC}}\Big(g\big(\overline{X}(T)\big);\tilde{k}_0\Big)-\mathrm{Var}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big)\big)\right|>\mathrm{Var}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big)\big)\epsilon\right)\\
&\leq P\Bigg(\left|\mathcal{V}\Big(g\big(\overline{X}_0(T)\big);\tilde{k}_0\Big)-\mathrm{Var}\big(g\big(\overline{X}_0(T)\big)\big)\right|+\sum_{\ell=1}^{L}2^L\left\lceil 2^{L+\ell(\bar{\gamma}-1)}\right\rceil^{-1}\\
&\quad\times\left|\mathcal{V}\Big(\Delta_\ell g\big(\overline{X}(T)\big);\tilde{k}_\ell\Big)-\mathrm{Var}\big(\Delta_\ell g\big(\overline{X}(T)\big)\big)\right|>\mathrm{Var}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big)\big)\epsilon\Bigg)\\
&\leq P\left(\left|\mathcal{V}\Big(g\big(\overline{X}_0(T)\big);\tilde{k}_0\Big)-\mathrm{Var}\big(g\big(\overline{X}_0(T)\big)\big)\right|>\frac{\mathrm{Var}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big)\big)\epsilon}{L+1}\right)\\
&+\sum_{\ell=1}^{L}P\left(2^{(1-\bar{\gamma})\ell}\left|\mathcal{V}\Big(\Delta_\ell g\big(\overline{X}(T)\big);\tilde{k}_\ell\Big)-\mathrm{Var}\big(\Delta_\ell g\big(\overline{X}(T)\big)\big)\right|>\frac{\mathrm{Var}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big)\big)\epsilon}{L+1}\right)
\end{aligned}
$$

From the fourth moment bound (4.7), Chebycheff's inequality and k-Statistics bounds on the variance of the sample variance, cf. [17], we get that

$$
\begin{aligned}
&P\left(\left|\mathcal{V}\Big(g\big(\overline{X}_0(T)\big);\tilde{k}_0\Big)-\mathrm{Var}\big(g\big(\overline{X}_0(T)\big)\big)\right|>\frac{\mathrm{Var}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big)\big)\epsilon}{L+1}\right)\\
&\qquad\leq\frac{C(L+1)^2}{\mathrm{Var}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big)\big)^2\epsilon^2\tilde{k}_0}.
\end{aligned}
$$

The equality $2^{(1-\bar{\gamma})\ell}=\frac{\rho_{low}(\mathrm{TOL}_{\mathrm{T},\ell})\mathrm{TOL}_{\mathrm{T},0}}{\rho_{low}(\mathrm{TOL}_{\mathrm{T},0})\mathrm{TOL}_{\mathrm{T},\ell}}$ combined with (4.7) further yields that

$$
\begin{aligned}
&P\left(2^{(1-\bar{\gamma})\ell}\left|\mathcal{V}\Big(\Delta_\ell g\big(\overline{X}(T)\big);\tilde{k}_\ell\Big)-\mathrm{Var}\big(\Delta_\ell g\big(\overline{X}(T)\big)\big)\right|>\frac{\mathrm{Var}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big)\big)\epsilon}{L+1}\right)\\
&\leq\frac{C(L+1)^2}{\mathrm{Var}_{\mathcal{MLC}}\big(g\big(\overline{X}(T)\big)\big)^2\epsilon^2\tilde{k}_\ell}.
\end{aligned}
$$

Since $\tilde{k}_0=2^{L+\lceil C_{\mathcal{MLC}}L\rceil+1}$, the definition of $\tilde{k}_\ell$ in (4.29) implies that $\tilde{k}_\ell\geq 2^{L+\lceil C_{\mathcal{MLC}}L\rceil+1+(\bar{\gamma}-1)\ell}$ for $\ell=1,2,\ldots,L$, with $\bar{\gamma}\geq 0$ denoting the lower error

density exponent in $\rho_{low}(\mathrm{TOL_T}) = \mathrm{TOL_T}^{\bar{\gamma}}$, cf. (1.24). Consequently,

$$
\begin{aligned}
P\left(\left|\frac{\mathcal{V}_{\mathcal{MC}}\Big(g\big(\overline{X}(T)\big);\tilde{k}_0\Big)}{\mathrm{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big)}-1\right|>\epsilon\right) &\leq \frac{C(L+1)^2}{\mathrm{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big)^2\epsilon^2\tilde{k}_0}\sum_{\ell=0}^{L}\frac{\tilde{k}_0}{\tilde{k}_\ell}\\
&\leq \frac{C(L+1)^2}{\mathrm{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big)^2\epsilon^2\tilde{k}_0}\sum_{\ell=0}^{L}2^{(1-\bar{\gamma})\ell}\\
&< \frac{C(L+1)^2}{2^{\lceil C_{\mathcal{MC}}L\rceil+\bar{\gamma}L}\mathrm{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big)^2\epsilon^2}
\end{aligned}
$$

which implies that for any $\epsilon > 0$,

$$
\begin{aligned}
&\lim_{\mathrm{TOL}\downarrow 0} P\left(\left|\frac{\mathcal{V}_{\mathcal{MC}}\Big(g\big(\overline{X}(T)\big);\tilde{k}_0\Big)}{\mathrm{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big)}-1\right|>\epsilon\right)\\
&\qquad < \lim_{\mathrm{TOL}\downarrow 0}\frac{C(L+1)^2}{2^{\lceil C_{\mathcal{MC}}L\rceil+\bar{\gamma}L}\mathrm{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big)^2\epsilon^2}=0.
\end{aligned}
$$

Since $M_0 \geq \tilde{k}_0$ by definition, we conclude that also (4.32) holds, i.e.

$$
\lim_{\mathrm{TOL}\downarrow 0} P\left(\left|\frac{\mathcal{V}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big);M_0\big)}{\mathrm{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big)}-1\right|>\epsilon\right)=0,
$$

for any $\epsilon > 0$. □

With the asymptotic bounds on $M_0$ we are ready to prove the main asymptotic accuracy result, Theorem 2. For the convenience of the reader, we first recall its formulation.

**Theorem 2** (Multilevel accuracy)**.** *Suppose that the modeling assumptions of Lemma 1 hold, that* (4.7) *holds, and that* $\mathrm{TOL_T} \leq \mathrm{TOL_S}$. *Then the adaptive MLMC algorithm with confidence parameter* $C_C > 0$ *and stochastic time steps* (2.20) *and* (2.21) *satisfies*

$$
\liminf_{\mathrm{TOL}\downarrow 0} P\left(\left|\mathrm{E}[g(X(T))]-\mathcal{A}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big);M_0\big)\right|\leq \mathrm{TOL}\right) \geq \int_{-C_C}^{C_C}\frac{e^{-x^2/2}}{\sqrt{2\pi}}dx. \tag{4.33}
$$

*Proof.* This proof is quite similar to the proof of Theorem 4 for the asymptotic accuracy in the single level setting, but for the sake of the differing details, a full proof is included in this setting also. For a given $\delta > 0$, we start by bounding the left-hand side of (4.33) by a product of the statistical error and the time

discretization error.

$$
\begin{aligned}
&\liminf_{\mathrm{TOL}\downarrow 0} P\left(\left|\mathrm{E}[g(X(T))] - \mathcal{A}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big); M_0\big)\right| \leq \mathrm{TOL}\right) \\
&\geq \liminf_{\mathrm{TOL}\downarrow 0} P\left(\left|\mathrm{E}\left[g(X(T)) - g\big(\overline{X}_L(T)\big)\right]\right|\right. \\
&\qquad\qquad \left. + \left|\mathrm{E}\left[g\big(\overline{X}_L(T)\big)\right] - \mathcal{A}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big); M_0\big)\right| \leq C_S\mathrm{TOL}_{\mathrm{T}} + \mathrm{TOL}_{\mathrm{S}}\right) \\
&\geq \liminf_{\mathrm{TOL}\downarrow 0} P\left(\left|\mathrm{E}\left[g(X(T)) - g\big(\overline{X}_L(T)\big)\right]\right| \leq (C_S + \delta)\mathrm{TOL}_{\mathrm{T}}\right. \\
&\qquad\qquad \left. \text{and } \left|\mathrm{E}\left[g\big(\overline{X}_L(T)\big)\right] - \mathcal{A}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big); M_0\big)\right| \leq (1-\delta)\mathrm{TOL}_{\mathrm{S}}\right) \\
&= \liminf_{\mathrm{TOL}\downarrow 0} P\left(\left|\mathrm{E}\left[g(X(T)) - g\big(\overline{X}_L(T)\big)\right]\right| \leq (C_S + \delta)\mathrm{TOL}_{\mathrm{T}}\right) \\
&\qquad\qquad \times P\left(\left|\mathrm{E}\left[g\big(\overline{X}_L(T)\big)\right] - \mathcal{A}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big); M_0\big)\right| \leq (1-\delta)\mathrm{TOL}_{\mathrm{S}}\right)
\end{aligned}
$$

**The time discretization error.** The proof of Theorem 3.4, p. 530 in [19] shows that the following bound is fulfilled

$$
\lim_{\mathrm{TOL}\downarrow 0} \frac{\left|\mathrm{E}\left[g(X(T)) - g\big(\overline{X}_L(T)\big)\right]\right|}{\mathrm{TOL}_{\mathrm{T},L}} \leq C_S.
$$

By construction $\mathrm{TOL}_{\mathrm{T},L} = \mathrm{TOL}_{\mathrm{T}}$, and this implies by the above that

$$
\liminf_{\mathrm{TOL}\downarrow 0} P\left(\left|\mathrm{E}\left[g(X(T)) - g\big(\overline{X}_L(T)\big)\right]\right| \leq (1+\delta)C_S\mathrm{TOL}_{\mathrm{T}}\right) = 1.
$$

**The statistical error.** From the above introduced $\delta > 0$, define the family of sets

$$
\Omega_\delta(\mathrm{TOL}_{\mathrm{S}}) = \left\{ k \in 2^{\mathbb{N}+L+\lceil C_{\mathcal{MC}} L\rceil} \;\middle|\; 1-\delta < \frac{k{\mathrm{TOL}_{\mathrm{S}}}^2}{\mathrm{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big)C_C^2} \leq 2+\delta \right\}, \tag{4.34}
$$

indexed by $\mathrm{TOL}_{\mathrm{S}} > 0$. Lemma 7 then implies that $\lim_{\mathrm{TOL}\downarrow 0} P(M_0 \in \Omega_\delta) = 1$. Recall further that for the MLMC algorithm, the number of samples $M_0$ is determined in the step prior to generating the output $\mathcal{A}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big); M_0\big)$, so that $M_0$ is independent from $\mathcal{A}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big); M_0\big)$. Using this independence property

and Fatou's Lemma, the statistical error is bounded from below as follows:

$$
\begin{aligned}
&\liminf_{\mathrm{TOL}\downarrow 0} P\left(\left|E\left[g\big(\overline{X}_L(T)\big)\right] - \mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big); M_0\big)\right| \le (1-\delta)\mathrm{TOL}_{\mathrm{S}}\right) \\
&= \liminf_{\mathrm{TOL}\downarrow 0} \sum_{k_0 \in 2^{\mathbb{N}+L+\lceil C_{\mathcal{ML}} L\rceil}} P\big(\left|\mathrm{E}\left[g\big(\overline{X}_L(T)\big)\right] - \mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big); k_0\big)\right| \le (1-\delta)\mathrm{TOL}_{\mathrm{S}}\big) P(M_0 = k_0) \\
&\ge \liminf_{\mathrm{TOL}\downarrow 0} \sum_{k_0 \in \Omega_\delta} P\big(\left|\mathrm{E}\left[g\big(\overline{X}_L(T)\big)\right] - \mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big); k_0\big)\right| \le (1-\delta)\mathrm{TOL}_{\mathrm{S}}\big) P(M_0 = k_0) \\
&\quad + \sum_{k_0 \in 2^{\mathbb{N}+L+\lceil C_{\mathcal{ML}} L\rceil} \setminus \Omega_\delta} \liminf_{\mathrm{TOL}\downarrow 0} P\big(\left|\mathrm{E}\left[g\big(\overline{X}_L(T)\big)\right] - \mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big); k_0\big)\right| \le (1-\delta)\mathrm{TOL}_{\mathrm{S}}\big) P(M_0 = k_0) \\
&\ge \liminf_{\mathrm{TOL}\downarrow 0} \sum_{k_0 \in \Omega_\delta} P\left(\sqrt{k_0}\frac{\left|\mathrm{E}\left[g\big(\overline{X}_L(T)\big)\right] - \mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big); k_0\big)\right|}{\sqrt{\mathrm{Var}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big)\big)}} \le (1-\delta)^{3/2} C_C\right) P(M_0 = k_0) \\
&\ge \int_{-(1-\delta)^{3/2}C_C}^{(1-\delta)^{3/2}C_C} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx.
\end{aligned}
\tag{4.35}
$$

The last inequality above follows from the application of Lindeberg-Feller's Central Limit Theorem (CLT) which is justified by Lemma 8 and the observation that $\mathrm{E}\big[\mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big); k_0\big)\big] = \mathrm{E}\big[g\big(\overline{X}_L(T)\big)\big]$. The reasoning leading to inequality (4.35) is valid for any $\delta > 0$, so the proof of (4.33) is finished. □

Next we derive the weak convergence CLT result for the multilevel estimator $\mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big); k_0\big)$ which is needed in Theorem 2.

**Lemma 8** (A CLT result)**.** *Suppose the assumptions of Lemma 1 and* (4.7) *hold and, for a given $\delta > 0$, let*

$$
k_0(\mathrm{TOL}_{\mathrm{S}}) := \min\left\{ k \in 2^{\mathbb{N}+L+\lceil C_{\mathcal{ML}} L\rceil} \,\middle|\, \frac{k{\mathrm{TOL}_{\mathrm{S}}}^2}{\mathrm{Var}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big)\big) C_C^2} > 1-\delta \right\},
$$

*in correspondence with the set defined in* (4.34)*. Then for any $z \in \mathbb{R}_+$, we have that*

$$
\lim_{\mathrm{TOL}\downarrow 0} P\left(\sqrt{k_0}\frac{\left|\mathrm{E}\left[\mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big); k_0\big)\right] - \mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big); k_0\big)\right|}{\sqrt{\mathrm{Var}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big)\big)}} \le z\right) = \int_{-z}^{z} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx. \tag{4.36}
$$

*Proof.* This Lemma will be proved by verifying that the assumptions of the Lindeberg-Feller CLT are fulfilled, cf. Theorem 6. Let us write

$$
\sqrt{k_0}\frac{\mathrm{E}\left[\mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big); k_0\big)\right] - \mathcal{A}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big); k_0\big)}{\sqrt{\mathrm{Var}_{\mathcal{ML}}\big(g\big(\overline{X}(T)\big)\big)}} = \sum_{i=1}^{K} Y_{K,i}
$$

where $K := \sum_{\ell=0}^{L} k_\ell$ and the elements of $Y_{K,i}$ are independent and defined by

$$Y_{K,i} := \begin{cases} \frac{\mathrm{E}\left[g(\overline{X}_0(T))\right] - g(\overline{X}_0(T;\omega_i))}{\sqrt{k_0}\,\sqrt{\mathrm{Var}_{\mathcal{ML}}\left(g\left(\overline{X}(T)\right)\right)}} & \text{for } i = 1, 2, \ldots, k_0, \\ \frac{\sqrt{\frac{k_0}{k_1}}\left(\mathrm{E}\left[\Delta_1 g(\overline{X}(T))\right] - \Delta_1 g(\overline{X}(T;\omega_i))\right)}{\sqrt{k_1}\,\sqrt{\mathrm{Var}_{\mathcal{ML}}\left(g\left(\overline{X}(T)\right)\right)}} & \text{for } i = k_0 + 1, \ldots, k_0 + k_1 \\ \vdots & \vdots \\ \frac{\sqrt{\frac{k_0}{k_L}}\left(\mathrm{E}\left[\Delta_L g\left(\overline{X}(T)\right)\right] - \Delta_L g(\overline{X}(T;\omega_i))\right)}{\sqrt{k_L}\,\sqrt{\mathrm{Var}_{\mathcal{ML}}\left(g\left(\overline{X}(T)\right)\right)}} & \text{for } i = k_{L-1} + 1, \ldots, K. \end{cases}$$

Then it follows that

$$\sum_{i=1}^{K} \mathrm{E}\left[Y_{K,i}^2\right] = \frac{\mathrm{Var}_{\mathcal{ML}}\left(g\left(\overline{X}(T)\right)\right)}{\mathrm{Var}_{\mathcal{ML}}\left(g\left(\overline{X}(T)\right)\right)} = 1, \quad \forall \mathrm{TOL} > 0,$$

so condition (a) of Theorem 6 is fulfilled. To verify that condition (b) of Theorem 6 is fulfilled, one must show that for any $\epsilon > 0$,

$$\limsup_{\mathrm{TOL} \to 0} \sum_{i=1}^{K} \mathrm{E}\left[Y_{K,i}^2 1_{|Y_{K,i}| > \epsilon}\right] = 0.$$

The definition of $k_\ell$, cf. (4.29), combined with the moment bound (4.7) implies that there exists a $C > 0$ such that

$$\mathrm{E}\left[\left(\frac{k_0}{k_\ell}\right)^2 \left|\Delta_\ell g\left(\overline{X}(T)\right) - \mathrm{E}\left[\Delta_\ell g\left(\overline{X}(T)\right)\right]\right|^4\right] \leq C, \quad \forall \ell \in \{1, 2, \ldots, L\}.$$

Using Chebycheff's inequality and the fact that $k_L \geq 2^{\lceil C_{\mathcal{ML}} L \rceil + \bar{\gamma} L + 1}$, cf. (4.29), we derive that

$$\begin{aligned} \sum_{i=1}^{K} \mathrm{E}\left[Y_{K,i}^2 1_{|Y_{K,i}| > \epsilon}\right] &\leq \sum_{i=1}^{K} \epsilon^{-2} \mathrm{E}\left[Y_{K,i}^4\right] \\ &= \frac{1}{\epsilon^2\,\mathrm{Var}_{\mathcal{ML}}\left(g\left(\overline{X}(T)\right)\right)^2} \left\{ \frac{1}{k_0} \mathrm{E}\left[\left|g(\overline{X}_0(T)) - \mathrm{E}\left[g\left(\overline{X}_0(T)\right)\right]\right|^4\right] \right. \\ &\qquad \left. + \sum_{\ell=1}^{L} \frac{1}{k_\ell} \mathrm{E}\left[\left(\frac{k_0}{k_\ell}\right)^2 \left|\Delta_\ell g\left(\overline{X}(T)\right) - \mathrm{E}\left[\Delta_\ell g\left(\overline{X}(T)\right)\right]\right|^4\right] \right\} \\ &\leq \frac{C}{\epsilon^2\,\mathrm{Var}_{\mathcal{ML}}\left(g\left(\overline{X}(T)\right)\right)^2} \sum_{\ell=0}^{L} k_\ell^{-1} \\ &\leq \frac{C\,L}{k_L\,\epsilon^2\,\mathrm{Var}_{\mathcal{ML}}\left(g\left(\overline{X}(T)\right)\right)^2} \to 0, \quad \text{as } \mathrm{TOL} \downarrow 0. \end{aligned}$$

This verifies that condition (b) is fulfilled. □

We conclude the analysis of the MLMC algorithm by estimating the work required to fulfill the accuracy estimate (4.1). We recall that WORK(TOL), defined in (4.2) by

$$\text{WORK(TOL)} = \sum_{\ell=0}^{L} \mathrm{E}[M_\ell]\mathrm{E}[N_\ell],$$

is an estimate of the average number of operations required in the generation of $\mathcal{A}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big); M_0\big)$ to approximate $\mathrm{E}[g(X(T))]$ with the prescribed confidence $C_C$ and accuracy TOL. First, let us derive an asymptotic bound for $\mathrm{E}[M_0]$.

**Lemma 9.** *Suppose the assumptions of Lemma 1 and* (4.7) *hold. Then the number of samples $M_0$ used at the base level of the MLMC algorithm approximation of $\mathrm{E}[g(X(T))]$ satisfies*

$$\limsup_{\text{TOL}\downarrow 0} \frac{\mathrm{E}[M_0]\text{TOL}_{\text{S}}^2}{\text{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big)C_C^2} \leq 2. \tag{4.37}$$

*Proof.* For given $\delta > 0$, define the deterministic function

$$\widetilde{M}_0(\text{TOL}) = \min\left\{ k \in 2^{\mathbb{N}+L+\lceil C_{\mathcal{MC}} L\rceil} \Big| \frac{k_0 \text{TOL}^2}{\text{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big)C_C^2} > 1+\delta \right\}$$

By the relation (4.16), the moment bound assumption (4.7), Hölder's inequality, and k-Statistics bounds on the variance of the sample variance, cf. [17], we derive that

$$\begin{aligned}
P(M_0 = 2\widetilde{M}_0) &\leq P\left( \frac{\mathcal{V}_{\mathcal{MC}}\Big(g\big(\overline{X}(T)\big); \widetilde{M}_0\Big)}{\text{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big)} > \widetilde{M}_0 \frac{\text{TOL}^2}{\text{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big)C_C^2} \right) \\
&\leq P\left( \frac{\mathcal{V}_{\mathcal{MC}}\Big(g\big(\overline{X}(T)\big); \widetilde{M}_0\Big)}{\text{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big)} > 1+\delta \right) \\
&\leq P\left( \mathcal{V}_{\mathcal{MC}}\Big(g\big(\overline{X}(T)\big); \widetilde{M}_0\Big) - \text{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big) > \delta \text{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big) \right) \\
&\leq \mathrm{E}\left[ \frac{\Big|\mathcal{V}_{\mathcal{MC}}\Big(g\big(\overline{X}(T)\big); \widetilde{M}_0\Big) - \text{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big)\Big|^2}{\delta^2 \text{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big)^2} \right] \\
&\leq \frac{\text{Var}\Big(\mathcal{V}\Big(g\big(\overline{X}_0(T)\big); \widetilde{M}_0\Big)\Big) + \sum_{\ell=1}^{L} \text{Var}\Big(\mathcal{V}\Big(\Delta_\ell g\big(\overline{X}(T)\big); \widetilde{M}_\ell\Big)\Big)}{\delta^2 \text{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big)^2} \\
&\leq \frac{CL}{\delta^2 \text{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big)^2 \widetilde{M}_L},
\end{aligned}$$

and for $\ell = 1, 2, \ldots$ that

$$
\begin{aligned}
&P(M_0 = 2^{\ell+1}\widetilde{M}_0) \\
&\quad \leq P\left(\mathcal{V}_{\mathcal{MC}}\Big(g\big(\overline{X}(T)\big); 2^{\ell}\widetilde{M}_0\Big) - \mathrm{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big) > 2^{\ell-1}\mathrm{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big)\right) \\
&\quad \leq \mathrm{E}\left[\frac{\left|\mathcal{V}_{\mathcal{MC}}\Big(g\big(\overline{X}(T)\big); 2^{\ell}\widetilde{M}_0\Big) - \mathrm{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big)\right|^2}{2^{2(\ell-1)}\mathrm{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big)^2}\right] \\
&\quad < \frac{CL}{2^{3\ell}\,\mathrm{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big)^2\widetilde{M}_L}.
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
\frac{\mathrm{E}[M_0]\mathrm{TOL}_{\mathrm{S}}^2}{\mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)C_C^2} &\leq \left[P(M_0 \leq \widetilde{M}_0) + \sum_{\ell=1}^{\infty} 2^{\ell} P(M_0 = 2^{\ell}\widetilde{M}_0)\right]\frac{\widetilde{M}_0\mathrm{TOL}_{\mathrm{S}}^2}{\mathrm{Var}\big(g\big(\overline{X}(T)\big)\big)C_C^2} \\
&\leq 2(1+\delta)\left[P(M_0 \leq \widetilde{M}_0) + P(M_0 = 2\widetilde{M}_0) + \sum_{\ell=1}^{\infty} 2^{\ell+1} P(M_0 = 2^{\ell+1}\widetilde{M}_0)\right] \\
&\leq 2(1+\delta)\left[P(M_0 \leq \widetilde{M}_0) + \frac{CL}{\delta^2\widetilde{M}_L} + \frac{CL}{\widetilde{M}_L}\sum_{\ell=1}^{\infty} 2^{-2\ell}\right].
\end{aligned}
$$

Taking limits in the above inequality leads to

$$
\limsup_{\mathrm{TOL}\downarrow 0} \frac{\mathrm{E}[M_0]\mathrm{TOL}_{\mathrm{S}}^2}{\mathrm{Var}_{\mathcal{MC}}\big(g\big(\overline{X}(T)\big)\big)C_C^2} \leq 2(1+\delta).
$$

Finally, observe that since the obtained inequality holds true for any $\delta > 0$, the proof is finished. □

An asymptotic bound on $\mathrm{E}[N_\ell]$ may be deduced from the single level result of Lemma 6. For the convenience of the reader we present the result of Lemma 6 in a way that is fitting for the multilevel setting.

**Lemma 10** (Multilevel asymptotical average number of time steps)**.** *Suppose that the assumptions of Lemma 1 hold. Then the final number of adaptive steps generated by the MLMC algorithm with stochastic time steps* (2.20) *and* (2.21) *and* $\mathrm{TOL}_{\mathrm{T},\ell} = 2^{-\ell}\mathrm{TOL}_{\mathrm{T},0}$ *satisfies*

$$
\limsup_{\ell\uparrow\infty} \mathrm{TOL}_{\mathrm{T},\ell}\ E[N_\ell] \leq \frac{4}{C_R}\left(\mathrm{E}\left[\int_0^T \sqrt{|\hat{\rho}(t)|}dt\right]\right)^2. \tag{4.38}
$$

With bounds for $\mathrm{E}[M_0]$ and $\mathrm{E}[N_\ell]$ in hand, we are ready to prove the main complexity theorem for the MLMC algorithm, Theorem 3. The proof is divided into three cases.

*Proof of case (I).* Lemma 10 implies that for any given $\delta > 0$, there exists an $\hat{L}(\delta)$ not depending on TOL such that

$$\mathrm{TOL}_{\mathrm{T},\ell}\ \mathrm{E}[N_\ell] \leq (1+\delta)\frac{4}{C_R}\ \left(\mathrm{E}\left[\int_0^T \sqrt{|\widehat{\rho}(\tau)|}d\tau\right]\right)^2, \quad \forall \ell \geq \hat{L}. \tag{4.39}$$

Furthermore, for $\rho_{low}(\mathrm{TOL}_{\mathrm{T}}) = \rho_{\min}$, $M_\ell$ as defined in (4.26) fulfills

$$\mathrm{E}[M_\ell] = 2^{-\ell}\mathrm{E}[M_0], \quad \forall \ell \in \{0,1,\ldots,L\}. \tag{4.40}$$

By inequality (4.39), equation (4.40), and the monotonic relation $N_\ell \leq N_{\ell+1}$, we derive that

$$\begin{aligned}
\sum_{\ell=0}^{L} \mathrm{E}[M_\ell]\mathrm{E}[N_\ell] &\leq \mathrm{E}\left[N_{\widehat{L}}\right]\mathrm{TOL}_{\mathrm{T},\widehat{L}} \sum_{\ell=0}^{\widehat{L}} \frac{\mathrm{E}[M_\ell]}{\mathrm{TOL}_{\mathrm{T},\widehat{L}}} + \sum_{\ell=\widehat{L}+1}^{L} \frac{E[M_\ell]}{\mathrm{TOL}_{\mathrm{T},\ell}} E[N_\ell]\mathrm{TOL}_{\mathrm{T},\ell} \\
&\leq \frac{4(1+\delta)\mathrm{E}[M_0]}{C_R\,\mathrm{TOL}_{\mathrm{T},0}} \left(\mathrm{E}\left[\int_0^T \sqrt{|\widehat{\rho}(\tau)|}d\tau\right]\right)^2 \left(2^{\widehat{L}} \sum_{\ell=0}^{\widehat{L}-1} 2^{-\ell} + \sum_{\ell=\widehat{L}}^{L} 1\right) \\
&\leq \frac{4(1+\delta)\mathrm{E}[M_0]}{C_R\,\mathrm{TOL}_{\mathrm{T},0}} \left(\mathrm{E}\left[\int_0^T \sqrt{|\widehat{\rho}(\tau)|}d\tau\right]\right)^2 \left(2^{\widehat{L}+1} + (L-\widehat{L})\right).
\end{aligned} \tag{4.41}$$

Recalling the definition $L = \lfloor \log_2(\mathrm{TOL}_{\mathrm{T,Max}}/\mathrm{TOL}_{\mathrm{T}}) \rfloor$ and that $\hat{L}$ is fixed, it follows that

$$\lim_{\mathrm{TOL}\downarrow 0} \frac{2^{\widehat{L}+1} + (L-\widehat{L})}{L} = 1.$$

Using (4.41) combined with Lemma 9 and recalling that $\mathrm{TOL}_{\mathrm{T},0} > \mathrm{TOL}_{\mathrm{T,Max}}/2$, we obtain the bound

$$\limsup_{\mathrm{TOL}\downarrow 0} \frac{\mathrm{WORK}(\mathrm{TOL})\,\mathrm{TOL}_{\mathrm{S}}^2}{\mathrm{Var}_{\mathcal{MC}}\left(g\left(\overline{X}(T)\right)\right)L} \leq \frac{16(1+\delta)\mathrm{E}[M_0]}{C_R\,\mathrm{TOL}_{\mathrm{T,Max}}} \left(\mathrm{E}\left[\int_0^T \sqrt{|\widehat{\rho}(\tau)|}d\tau\right]\right)^2. \tag{4.42}$$

We observe that $\mathrm{WORK}(\mathrm{TOL}) = \mathcal{O}\left(\mathrm{TOL}_{\mathrm{S}}^{-2} L \mathrm{Var}_{\mathcal{MC}}\left(g\left(\overline{X}(T)\right)\right)\right)$, and to obtain a bound on more explicit form, we note that by the assumption (4.7) on $L^p$ convergence, there exists a $C_G > 0$ such that

$$\limsup_{\ell\uparrow\infty} \frac{\rho_{low}(\mathrm{TOL}_{\mathrm{T},\ell})}{\mathrm{TOL}_{\mathrm{T},\ell}} \mathrm{E}\left[\left|\Delta_\ell g\left(\overline{X}(T)\right)\right|^2\right] \leq C_G. \tag{4.43}$$

See for instance Remark 4 for a discussion on how to estimate $C_G$. Inequality (4.43) further implies that

$$\limsup_{\mathrm{TOL}\downarrow 0} \frac{\mathrm{Var}_{\mathcal{MC}}\left(g\left(\overline{X}(T)\right)\right)}{L} \leq C_G. \tag{4.44}$$

To approximately minimize the complexity we introduce the splitting choice

$$\mathrm{TOL_S} = \frac{\log(\mathrm{TOL}^{-1})}{\log(\mathrm{TOL}^{-1}) + \log(\log(\mathrm{TOL}^{-1}))}\mathrm{TOL}, \quad \text{and}$$
$$\mathrm{TOL_T} = \frac{\log(\log(\mathrm{TOL}^{-1}))}{(\log(\mathrm{TOL}^{-1}) + \log(\log(\mathrm{TOL}^{-1})))C_S}\mathrm{TOL},$$

which we see fulfills the restrictions $C_S\mathrm{TOL_T} + \mathrm{TOL_S} = \mathrm{TOL}$ and $\mathrm{TOL_T} \leq \mathrm{TOL_S}$. Combining (4.43) with the above splitting choice in inequality (4.42), and noting that this bounding procedure is valid for any $\delta > 0$ leads to (4.4). □

*Proof of case (II).* Recall that $M_\ell$ as defined in (4.26) fulfills

$$\mathrm{E}[M_\ell] \leq (2^{\ell(\bar{\gamma}-1)} + 2^{-L})\mathrm{E}[M_0], \quad \forall \ell \in \{0, 1, \ldots, L\}.$$

By this property, intequality (4.39) and recalling that by construction $\mathrm{TOL_{T,0}} > \mathrm{TOL_{T,Max}}/2$, we derive the following

$$\begin{aligned}
&\sum_{\ell=0}^{L} \mathrm{E}[M_\ell]\mathrm{E}[N_\ell] \leq \mathrm{E}\left[N_{\widehat{L}}\right]\mathrm{TOL}_{\mathrm{T},\widehat{L}} \sum_{\ell=0}^{\widehat{L}} \frac{\mathrm{E}[M_\ell]}{\mathrm{TOL}_{\mathrm{T},\widehat{L}}} + \sum_{\ell=\widehat{L}+1}^{L} \frac{E[M_\ell]}{\mathrm{TOL}_{\mathrm{T},\ell}} E[N_\ell]\mathrm{TOL}_{\mathrm{T},\ell} \\
&\leq \frac{(1+\delta)4\mathrm{E}[M_0]}{C_R\,\mathrm{TOL}_{\mathrm{T},0}} \left(\mathrm{E}\left[\int_0^T \sqrt{|\widehat{\rho}(\tau)|}d\tau\right]\right)^2 \\
&\quad \times \left(2^{\widehat{L}} \sum_{\ell=0}^{\widehat{L}-1} (2^{\ell(\bar{\gamma}-1)} + 2^{-L}) + \sum_{\ell=\widehat{L}}^{L} (2^{\ell\bar{\gamma}} + 2^{-L+\ell})\right) \\
&\leq \frac{(1+\delta)8\mathrm{E}[M_0]}{C_R\,\mathrm{TOL}_{\mathrm{T,Max}}} \left(\mathrm{E}\left[\int_0^T \sqrt{|\widehat{\rho}(\tau)|}d\tau\right]\right)^2 \left(\frac{2^{\widehat{L}}}{1-2^{\bar{\gamma}-1}} + \widehat{L}2^{\widehat{L}-L} + \frac{2^{(L+1)\bar{\gamma}}}{\log(2^{\bar{\gamma}})} + 2\right).
\end{aligned} \tag{4.45}$$

By noting that $\widehat{L}(\delta)$ is fixed, we have

$$\lim_{\mathrm{TOL}\downarrow 0} \frac{\frac{2^{\widehat{L}}}{1-2^{\bar{\gamma}-1}} + \widehat{L}2^{\widehat{L}-L} + \frac{2^{(L+1)\bar{\gamma}}}{\log(2^{\bar{\gamma}})} + 2}{2^{\bar{\gamma}L}} = \frac{2^{\bar{\gamma}}}{\bar{\gamma}\log(2)},$$

and by Lemma 9 and (4.44), we derive the bound

$$\begin{aligned}
&\limsup_{\mathrm{TOL}\downarrow 0} \frac{\mathrm{WORK}(\mathrm{TOL})\,\mathrm{TOL_S^2}}{2^{\bar{\gamma}L}L} \\
&\qquad \leq (1+\delta)\frac{2^{4+\bar{\gamma}}\,C_C^2\,C_G}{\bar{\gamma}\,\log(2)\,\mathrm{TOL}_{\mathrm{T,Max}}\,C_R} \left(\mathrm{E}\left[\int_0^T \sqrt{|\widehat{\rho}(\tau)|}d\tau\right]\right)^2.
\end{aligned}$$

Since $2^{-L} \geq \mathrm{TOL_T}/\mathrm{TOL_{T,Max}}$ by construction, we further obtain

$$\limsup_{\mathrm{TOL}\downarrow 0} \frac{\mathrm{WORK(TOL)}\,\mathrm{TOL_S^2}\,\mathrm{TOL_T^{\bar{\gamma}}}}{L} \leq (1+\delta)\frac{2^{4+\bar{\gamma}}\,C_C^2\,C_G}{\bar{\gamma}\,\log(2)\,(\mathrm{TOL_{T,Max}})^{1-\bar{\gamma}}\,C_R}\left(\mathrm{E}\left[\int_0^T \sqrt{|\widehat{\rho}(\tau)|}d\tau\right]\right)^2. \tag{4.46}$$

To approximately minimize the complexity, we introduce the splitting choice

$$\mathrm{TOL_S} = \frac{2}{2+\bar{\gamma}}\mathrm{TOL}, \qquad \text{and,} \qquad \mathrm{TOL_T} = \frac{\bar{\gamma}}{(2+\bar{\gamma})C_S}\mathrm{TOL},$$

which fulfills the constraints $C_S\mathrm{TOL_T} + \mathrm{TOL_S} = \mathrm{TOL}$ and $\mathrm{TOL_T} \leq \mathrm{TOL_S}$. Using the splitting choice in inequality (4.46) leads to (4.4) when noting that the bounding procedure is valid for any $\delta > 0$. □

*Proof of case (III).* First, we note that the conditions $\bar{\gamma} \to 0$ and $L\bar{\gamma} \to \infty$ as $\mathrm{TOL} \downarrow 0$ yields a consistent lower error density, since it leads to

$$\rho_{low}(\mathrm{TOL_T}) = \mathrm{TOL_T^{\bar{\gamma}}} = \mathcal{O}\left(2^{-L\bar{\gamma}}\right),$$

which implies that $\rho_{low}(\mathrm{TOL_T}) \to 0$ as $\mathrm{TOL} \downarrow 0$.

For proving case (III), recall inequality (4.45) from the proof of case (II) which is valid in the present setting as well, namely

$$\sum_{\ell=0}^{L} \mathrm{E}[M_\ell]\mathrm{E}[N_\ell] \leq \frac{(1+\delta)8\mathrm{E}[M_0]}{C_R\,\mathrm{TOL_{T,Max}}}\left(\mathrm{E}\left[\int_0^T \sqrt{|\widehat{\rho}(\tau)|}d\tau\right]\right)^2 \times \left(\frac{2^{\widehat{L}}}{1-2^{\bar{\gamma}-1}} + \widehat{L}2^{\widehat{L}-L} + \frac{2^{(L+1)\bar{\gamma}}}{\log(2^{\bar{\gamma}})} + 2\right).$$

The asymptotic conditions on $\bar{\gamma}$ imply that

$$\lim_{\mathrm{TOL}\downarrow 0} \frac{\bar{\gamma}}{2^{\bar{\gamma}L}}\left(\frac{2^{\widehat{L}}}{1-2^{\bar{\gamma}-1}} + \widehat{L}2^{\widehat{L}-L} + \frac{2^{(L+1)\bar{\gamma}}}{\log(2^{\bar{\gamma}})} + 2\right) = \frac{1}{\log(2)}.$$

This property, Lemma 9 and (4.44), yield

$$\limsup_{\mathrm{TOL}\downarrow 0} \frac{\mathrm{WORK(TOL)}\,\mathrm{TOL_S^2}\,\bar{\gamma}}{L\,2^{\bar{\gamma}L}} \leq (1+\delta)\frac{16\,C_C^2\,C_G}{\log(2)\,\mathrm{TOL_{T,Max}}\,C_R}\left(\mathrm{E}\left[\int_0^T \sqrt{|\widehat{\rho}(\tau)|}d\tau\right]\right)^2. \tag{4.47}$$

To approximately minimize the complexity, we introduce the splitting choice

$$\mathrm{TOL_S} = \frac{2}{2+\bar{\gamma}(\mathrm{TOL})}\mathrm{TOL} \qquad \text{and} \qquad \mathrm{TOL_T} = \frac{\bar{\gamma}(\mathrm{TOL})}{(2+\bar{\gamma}(\mathrm{TOL}))C_S}\mathrm{TOL}.$$

Inserting the splitting into (4.47) and noting that this argument is valid for all $\delta > 0$ leads to (4.6). □

**Remark 4** (Particular estimate for the constant $C_G$)**.** *It is possible to estimate the constant $C_G$. For instance, in the particular case when the exact error density is bounded away from zero, i.e. there exist a constant $\rho_{min}$ such that $\hat{\rho} > \rho_{min} > 0$ a.s. and, considering the equation*

$$\begin{aligned} dX(t) &= b(X(t))dW(t),\ t > 0 \\ X(0) &= X_0, \end{aligned}$$

*we have*

$$C_G \leq C_S \,\mathrm{E}\left[ \left\| \frac{(b'b)^2(X(t))(\varphi)^2(t)}{\hat{\rho}(t)} \right\|_{L^\infty([0,T])} \right].$$

*Here $\varphi(t) = g'(X(T))\frac{X'(T)}{X'(t)}$ and the first variation $X'(s)$ solves, for $s > 0$, the linear equation*

$$dX'(s) = b'(X(s))X'(s)dW(s),$$

*with initial condition $X'(0) = 1$. The constant $C_S$ is the parameter in the stopping condition* (2.20)*.*

**Remark 5** (Jump Diffusions)**.** *It is possible to extend these results of adaptive multilevel weak approximation for diffusions to the case of jump diffusions with time dependent jump measure analyzed in [22].*

# 5 Conclusions

In this paper we have presented and analyzed an adaptive multilevel Monte Carlo algorithm, where the multilevel simulations are performed on adaptively generated mesh hierarchies based on computable a posteriori weak error estimates. The theoretical analysis of the adaptive algorithm showed that the algorithm stops after a finite number of steps, and proceeded to show accuracy and efficiency results under natural assumptions in Theorems 2 and 3. In particular, Theorem 2 states that the probability of the weak error being bounded by the specified tolerance TOL is asymptotically bounded by any desired probability through the confidence parameter. Theorem 3 states computational complexity results where the involved constants are explicitly given in terms of algorithm parameters and problem properties. It shows that the $L^{1/2}$-quasi norm of the error density appears as a multiplicative constant in the complexity bounds, instead of the larger $L^1$-norm of the same error density that would appear using a uniform time step MLMC algorithm; the difference between these two factors can be arbitrarily large even in problems with smooth coefficients where they are both finite. Disregarding the constants the result shows that, depending on assumptions on the limit error density and the lower bound on the computed error density used by the adaptive algorithm, the complexity can be either the

same as or nearly the same as the complexity uniform MLMC has in cases where the order of strong convergence of the Euler-Maruyama method is 1/2.

Numerical results for scalar SDEs confirmed the theoretical analysis. For the two problems with reduced weak convergence order a simple single level Monte Carlo method has complexity $\mathcal{O}(\mathrm{TOL}^{-4})$ while the adaptive MLMC method has the improved complexity $\mathcal{O}(\mathrm{TOL}^{-2}\log_2(\mathrm{TOL}_0/\mathrm{TOL})^2)$. The use of advanced Monte Carlo methods such as the adaptive MLMC algorithm presented in this paper is most attractive for SDEs in higher dimension, where the corresponding PDE-based computational techniques are not competitive. Numerical experiments using adaptive MLMC is ongoing work and will be presented in a later report; this also makes for an interesting comparison to uniform MLMC method for barrier problems, since it is not clear that the order of strong convergence of the Euler-Maruyama method will be $(1-\delta)/2$, for any positive $\delta$, in that case. The fact that computational complexity of uniform multilevel Monte Carlo, disregarding constants, depends on the strong convergence indicates that adaptive mesh refinements based on strong error estimates can also be used to improve the computational efficiency; such methods are also subjects of ongoing research.

In this paper the adaptive algorithms were presented with global error control in the quantity of interest, starting from a given coarse mesh. Alternatively we can use local error estimates to guide adaptive time step control in the computation of the forward problem. This approach can be used on its own when global error control is deemed unnecessary or too computationally expensive, but it can also be used together with the global error control in situations with stiff SDEs where any given initial mesh can be too coarse depending on the realization. This is particularly relevant for MLMC simulations where stability issues in the computations on the coarsest level can destroy the results of the whole multilevel simulation, as was pointed out by Hutzenthaler, Jentzen, and Kloeden in [13].

# A Theorems

**Theorem 6** (Lindeberg-Feller Theorem [4][p. 114])**.** *For each $n$, let $X_{n,m}$, $1 \leq m \leq n$, be independent random variables with $E[X_{n,m}] = 0$. Suppose*

*(a)*

$$\sum_{m=1}^{n} E[X_{n,m}^2] \to \sigma^2 > 0, \text{ and}$$

*(b) for all $\delta > 0$,*

$$\lim_{n\to\infty} \sum_{m=1}^{n} E[X_{n,m}^2 1_{|X_{n,m}|>\delta}] = 0.$$

*Then the Central Limit Theorem holds, i.e., the random variable*

$$S_n := \sum_{m=1}^{n} X_{n,m} \rightharpoonup \sigma\Xi, \text{ as } n \to \infty,$$

*where* $\Xi$ *is a standard normal distributed random variable.*

**Acknowledgments.**
This work was partially supported by the Dahlquist fellowship at the Royal Institute of Technology in Stockholm, Sweden, the department of Scientific Computing in Florida State University, the University of Austin Subcontract (Project Number 024550, Center for Predictive Computational Science), the VR project "Effektiva numeriska metoder för stokastiska differentialekvationer med tillämpningar", and King Abdullah University of Science and Technology (KAUST). The authors would like to thank Mike Giles, and also two anonymous reviewers, for useful comments.

# References

[1] F. Kh. Abdullaev, J. C. Bronski, and G. Papanicolaou. Soliton perturbations and the random Kepler problem. *Phys. D*, 135(3-4):369–386, 2000.

[2] Eric Cancès, Frédéric Legoll, and Gabriel Stoltz. Theoretical and numerical comparison of some sampling methods for molecular dynamics. *M2AN Math. Model. Numer. Anal.*, 41(2):351–389, 2007.

[3] Y. S. Chow and Herbert Robbins. On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *The Annals of Mathematical Statistics*, 36(2):pp. 457–462, 1965.

[4] Richard Durrett. *Probability: theory and examples.* Duxbury Press, Belmont, CA, second edition, 1996.

[5] Anna Dzougoutov, Kyoung-Sook Moon, Erik von Schwerin, Anders Szepessy, and Raúl Tempone. Adaptive Monte Carlo algorithms for stopped diffusion. In *Multiscale methods in science and engineering*, volume 44 of *Lect. Notes Comput. Sci. Eng.*, pages 59–88. Springer, Berlin, 2005.

[6] Michael B. Giles. Improved multilevel monte carlo convergence using the milstein scheme. *Monte Carlo and Quasi-Monte Carlo Methods 2006*, pages 343–358.

[7] Michael B. Giles. Multilevel Monte Carlo path simulation. *Oper. Res.*, 56(3):607–617, 2008.

[8] Michael B. Giles, Desmond J. Higham, and Xuerong Mao. Analysing multilevel Monte Carlo for options with non-globally Lipschitz payoff. *Finance and Stochastics*, 13:403–413, 2009.

[9] Paul Glasserman. *Monte Carlo methods in financial engineering*, volume 53 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 2004. Stochastic Modelling and Applied Probability.

[10] Emmanuel Gobet. Weak approximation of killed diffusion using Euler schemes. *Stochastic Process. Appl.*, 87(2):167–197, 2000.

[11] S. Heinrich. Monte Carlo complexity of global solution of integral equations. *J. Complexity*, 14(2):151–175, 1998.

[12] S. Heinrich and E. Sindambiwe. Monte Carlo complexity of parametric integration. *J. Complexity*, 15(3):317–341, 1999. Dagstuhl Seminar on Algorithms and Complexity for Continuous Problems (1998).

[13] Martin Hutzenthaler, Arnulf Jentzen, and Peter E. Kloeden. Divergence of the multilevel monte carlo method. *arXiv/1105.0226v1*, 2011.

[14] E. Jouini, J. Cvitanić, and Marek Musiela, editors. *Option pricing, interest rates and risk management*. Handbooks in Mathematical Finance. Cambridge University Press, Cambridge, 2001.

[15] Ioannis Karatzas and Steven E. Shreve. *Brownian motion and stochastic calculus*, volume 113 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1991.

[16] Ahmed Kebaier. Statistical Romberg extrapolation: A new variance reduction method and applications to option pricing. *Ann. Appl. Probab.*, 15(4):2681–2705, 2005.

[17] E. S. Keeping. *Introduction to statistical inference*. D. Van Nostrand Co., Inc., Princeton, N.J.-Toronto-London-New York, 1962.

[18] Andrew J. Majda, Ilya Timofeyev, and Eric Vanden Eijnden. A mathematical framework for stochastic climate models. *Comm. Pure Appl. Math.*, 54(8):891–974, 2001.

[19] Kyoung-Sook Moon, Anders Szepessy, Raúl Tempone, and Georgios E. Zouraris. Convergence rates for adaptive weak approximation of stochastic differential equations. *Stoch. Anal. Appl.*, 23(3):511–558, 2005.

[20] Kyoung-Sook Moon, Erik von Schwerin, Anders Szepessy, and Raúl Tempone. An adaptive algorithm for ordinary, stochastic and partial differential equations. In *Recent advances in adaptive computation*, volume 383 of *Contemp. Math.*, pages 325–343. Amer. Math. Soc., Providence, RI, 2005.

[21] E. Mordecki, A. Szepessy, R. Tempone, and G. E. Zouraris. Adaptive weak approximation of diffusions with jumps. *SIAM J. Numer. Anal.*, 46(4):1732–1768, 2008.

[22] E. Mordecki, A. Szepessy, R. Tempone, and G. E. Zouraris. Adaptive weak approximation of diffusions with jumps. *SIAM J. Numer. Anal.*, 46(4):1732–1768, 2008.

[23] Bernt Øksendal. *Stochastic differential equations*. Universitext. Springer-Verlag, Berlin, fifth edition, 1998. An introduction with applications.

[24] Anders Szepessy, Raúl Tempone, and Georgios E. Zouraris. Adaptive weak approximation of stochastic differential equations. *Comm. Pure Appl. Math.*, 54(10):1169–1214, 2001.

# Paper III

# ON NON-ASYMPTOTIC OPTIMAL STOPPING CRITERIA IN MONTE CARLO SIMULATIONS

CHRISTIAN BAYER, HÅKON HOEL, ERIK VON SCHWERIN, AND RAÚL TEMPONE

Abstract. We consider the setting of estimating the mean of a random variable by a sequential stopping rule Monte Carlo (MC) method. The performance of a typical second moment based sequential stopping rule MC method is shown to be unreliable in such settings both by numerical examples and through analysis. By analysis and approximations, we construct a higher moment based stopping rule which is shown in numerical examples to perform more reliably and only slightly less efficiently than the second moment based stopping rule.

## 1. Introduction

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

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

We consider the objective of choosing $M$ sufficiently large so that the error probability satisfies

$$\mathrm{P}\big(\big|\overline{X}_M - \mu\big| > \mathrm{TOL}\big) \leq \delta, \tag{1}$$

for some fixed small constants $\mathrm{TOL} > 0$ and $\delta > 0$. Clearly, $\mathrm{P}\big(\big|\overline{X}_M - \mu\big| > \mathrm{TOL}\big)$ 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 derive theory or algorithms that will give upper bounds on $M$ satisfying (1) that are not far too large. When a-priori information about the distribution of $X$ is available, for example if $X$ is a bounded r.v. with an explicitly given bound, it is possible to derive good theoretical upper bounds for $M$ using Hoeffding type inequalities, cf. Hoeffding [5]. In the general case when no or little information of the distribution is given, little theory is however known, and the typical way of estimating $\mathrm{E}[X]$ using a sufficiently large number of samples $M$ is through a sequential stopping rule. Below we give the general structure of the class of sequential stopping rules we have in mind.

(I) Generate a batch of $M$ i.i.d. samples $X_1, X_2, \ldots, X_M$.

(II) Infer distributive properties of $\overline{X}_M$ from the generated batch of samples through higher order sample moments, e.g., the sample mean and the sample variance.

(III) Based on the sample moments, estimate the error probability. When, based on the estimated probability, (1) is violated, increase the number of samples $M$ and return to step (I).
Else, break and accept $M$.

2010 *Mathematics Subject Classification.* Primary 65C05; Secondary 62L12, 62L15.

*Key words and phrases.* Monte Carlo methods, optimal stopping, sequential stopping rules, non-asymptotic.

**Algorithm 1** Sample Variance Based Stopping Rule

**Input:** Number of samples $M_0$, accuracy TOL, confidence $\delta$, the cumulative distribution function of the standard normal distributed r.v. $\Phi(x)$.
**Output:** $\overline{X}_{M_{\overline{n}}}$.

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

$$\overline{\sigma}^2_{M_k} := \frac{1}{M_k - 1} \sum_{i=1}^{M_k} (X_i - \overline{X}_{M_k})^2. \tag{2}$$

**while** $2\Big(1 - \Phi(\sqrt{M_k}\,\mathrm{TOL}\,/\overline{\sigma}_{M_k})\Big) > \delta$ **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 $\overline{\sigma}^2_{M_k}$ as given in (2).
**end while**
Set $M_{\overline{n}} = M_k$, generate samples $\{X_i\}_{i=1}^{M_{\overline{n}}}$ and compute the output sample mean $\overline{X}_{M_{\overline{n}}}$. (See Section 2 for a motivation of the choice of the stopping criterion in the while loop above.)

Certainly the most natural and important representative of this class of algorithms is given in Algorithm 1. The algorithm estimates the error probability by appealing to the Central Limit Theorem (CLT). Consequently, it only relies on the sample variance in addition to the sample mean. In particular, the algorithm only requires mild additional assumptions on $X$, namely square integrability.

In the literature, various second moment based sequential stopping rules have been introduced to estimate the steady-state mean of stochastic processes, see for example Law and Kelton [7, 8] for comparisons of the performance of different stopping rules and Bratley, Bennet, and Fox [1] for an overview. Second moment based sequential stopping rules generally tend to perform well in the asymptotic regime when $\mathrm{TOL} \to 0$. In fact, Chow and Robbins [2] proved that under very loose restrictions, second moment based sequential stopping rules such as Algorithm 1 are asymptotically consistent, meaning that for a fixed $\delta$,

$$\lim_{\mathrm{TOL}\to 0} \mathrm{P}\big(\big|\overline{X}_M - \mu\big| > \mathrm{TOL}\big) = \delta,$$

and in Glynn and Whitt [4] the consistency property is proven to hold for such stopping rules applied to more general stochastic processes. The performance for second moment based stopping rules in the non-asymptotic regime – when TOL and $\delta$ are fixed values – is however not as well understood. This is unsatisfactory, as in applications this is precisely the interesting regime, in particular since very often we have $\mathrm{TOL} \gg \delta$. While consistency is clearly a re-assuring property in any case, in many situations one is in dire need of quantitative estimates of the error probability in the non-asymptotic regime, for instance when one tries to optimize the computational cost needed to meet a certain accuracy target using an adaptive algorithm. We could not find such a quantitative, non-asymptotic analysis of algorithms like Algorithm 1 in the literature.

In this note we demonstrate by numerical examples that second moment based stopping rules can fail convincingly in the non-asymptotic regime, especially when the underlying distribution $X$ is heavy-tailed, see Section 2. We proceed by giving an error analysis of Algorithm 1 specifically in the non-asymptotic regime. We note a-priori that there are two obvious approximation errors in the underlying assumptions of Algorithm 1:

(I) The algorithm appeals to the CLT to approximate the tail probabilities for $\overline{X}_M$ even though $M$ is finite.
(II) In doing so, it uses the sample variance $\overline{\sigma}^2_M$ instead of the true variance $\sigma^2$.

To get a hold on the error probability (1) despite the fact that the distribution of the sample mean $\overline{X}_M$ is unknown, we again appeal to the central limit theorem, but we adjust the estimate by adding a Berry-Esseen type term, which extends the validity of the estimate to the non-asymptotic case, thereby dealing with the first approximation error. As the error probability (1) is a tail probability for the distribution of the sample mean and the Berry-Esseen theorem itself is rather aimed at being sharp at the center of the distribution, we appeal to non-uniform versions of the Berry-Esseen theorem, see Theorem 1.1 and Corollary 1.2 below. However, both intuition and numerical tests suggest that the approximation of the tail probabilities by the non-uniform Berry-Esseen theorem is far too pessimistic at least when the second approximation error is small, i.e., when the computed sample variance is actually close to the true variance. In this case, we adjust the normal distribution by a less stringent extra term, which is obtained from an Edgeworth expansion of the distribution function of the sample mean $\overline{X}_M$, c.f. Feller [3].[1]

Having identified possible origins of failure of Algorithm 1, we propose an improvement of Algorithm 1. However, this variant requires third and fourth sample moments, see Section 4. Finally, in Section 5, we test the new algorithm numerically. We find that the new stopping Algorithm 2 indeed satisfies the desired confidence level $\delta$ on the error probability (1) even when $\delta \ll \mathrm{TOL}$.

As already discussed above, we need to approximate the unknown distribution of a sample mean in a general, non-asymptotic regime. The uniform and non-uniform Berry-Esseen theorems provide quantitative bounds for the difference between the true distribution of the sample mean and its asymptotic limit, namely the normal distribution. The following classical theorem can be found, for instance, in Petrov [10].

**Theorem 1.1** (Uniform and Non-Uniform Berry-Esseen)**.** *Suppose $X_1, X_2, \ldots$ are i.i.d. r.v. with $E[X] = 0$, $\sigma^2 = Var(X)$ and $\beta = \frac{E[|X|^3]}{\sigma^3} < \infty$. Then, for a positive constant $C_0$, the following uniform bound*

$$\left| P\left( \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} X_i < x \right) - \Phi(x) \right| \leq C_0 \frac{\beta}{\sqrt{n}}$$

*holds. For another positive constant $C_1$, the following non-uniform bound holds*

$$\left| P\left( \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} X_i < x \right) - \Phi(x) \right| \leq C_1 \frac{\beta}{\sqrt{n}(1 + |x|^3)}.$$

Up to our knowledge, the best upper bounds presently known for the Berry-Esseen constants are $C_0 < 0.4785$, cf. Tyurin [11], and $C_1 < C_0 + 8(1+e^1) < 30.2338$,

[1]Note that here we are introducing a gap in the analysis: the estimate based on the non-uniform Berry-Esseen theorem is reliable in the sense that it always leads to an upper bound of the error probability (1). For the Edgeworth expansion, however, there might be situations when the true error probability is underestimated, and, consequently, the accuracy target might still be missed. Numerical evidence, however, suggests that the estimate obtained from solely relying on the non-uniform Berry-Esseen theorem is usually by orders of magnitude too pessimistic. Apart from intrinsic reasons, one reason might be that the constants known in the non-uniform Berry-Esseen theorems might be far from being optimal. In the end, we think that the above compromise between Berry-Esseen type estimations and estimations based on the Edgeworth expansion might be a good compromise which retains the goal of reliably meeting the accuracy target – except maybe for very extreme situations – while keeping a certain level of efficiency. We note, however, that it is also possible to construct even more conservative stopping rules which are only based on the Berry-Esseen theorem.

cf. Michel [9]. For the purpose of this paper, it will be useful to combine the uniform and non-uniform Berry-Esseen bound as follows.

**Corollary 1.2** (Berry-Esseen)**.** *Suppose $X_1, X_2, \ldots$ are i.i.d. r.v. with $E[X] = 0$, $\sigma^2 = Var(X)$ and $\beta = E\big[|X|^3\big]/\sigma^3 < \infty$. Then*

$$\left| P\left( \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} X_i < x \right) - \Phi(x) \right| \leq C_{\mathrm{BE}}(x)\, \frac{\beta}{\sqrt{n}}$$

*where the bound function $C_{\mathrm{BE}} : \mathbb{R} \to [0, C_0]$ is defined by*

$$C_{\mathrm{BE}}(x) := \min\left( C_0, \frac{C_1}{(1+|x|)^3} \right).$$

In the asymptotic regime, the distribution of $\mathrm{P}\Big(\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} X_i < x\Big)$ can be expressed by so called Edgeworth expansions. Here we present the one-term Edgeworth expansion.

**Theorem 1.3** (Edgeworth expansion, cf. Feller [3])**.** *Suppose $X_1, X_2, \ldots$ are i.i.d. r.v. with a distribution which is not a lattice distribution and $E[X] = 0$, $\sigma^2 = Var(X)$ and $E\big[X^3\big] < \infty$. Then*

$$P\left( \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} X_i < x \right) = \Phi(x) + \frac{(x^2-1)e^{-x^2/2}}{6\sqrt{2\pi n}} \frac{E\big[X^3\big]}{\sigma^3} + o\Big(n^{-1/2}\Big),$$

*uniformly for $x \in \mathbb{R}$.*

## 2. Stopping rule failures

Suppose we seek to estimate $\mu = \mathrm{E}[X]$ using Monte Carlo simulation and we actually *do know* the variance $\sigma^2 = \mathrm{Var}\,(X)$. As before, our objective is to achieve $\mathrm{P}\big(\big|\overline{X}_M - \mu\big| > \mathrm{TOL}\big) \leq \delta$, for some fixed, small constants $\mathrm{TOL}, \delta > 0$. The CLT motivates the stopping rule

$$M = \frac{C_{\mathrm{CLT}}^2 \sigma^2}{\mathrm{TOL}^2}, \quad C_{\mathrm{CLT}} := \Phi^{-1}\left( \frac{2-\delta}{2} \right), \tag{3}$$

which would exactly fulfill our objective (1) in the asymptotic regime $M \to \infty$. Of course, this conflicts with our choice (3) for $M$, since we treat $\delta$ and TOL as finite constants. However, we can still estimate the probability in (1) using Corollary 1.2 and obtain

$$\begin{aligned}
\mathrm{P}\big(\big|\overline{X}_M - \mu\big| > \mathrm{TOL}\big) &= \mathrm{P}\left( \sqrt{M}\, \frac{\big|\overline{X}_M - \mu\big|}{\sigma} > \frac{\sqrt{M}\,\mathrm{TOL}}{\sigma} \right) \\
&\leq 2\left( 1 - \Phi\left( \frac{\sqrt{M}\,\mathrm{TOL}}{\sigma} \right) \right) + 2C_{\mathrm{BE}}\left( \frac{\sqrt{M}\,\mathrm{TOL}}{\sigma} \right) \frac{\beta}{\sqrt{M}} \qquad (4) \\
&= 2(1 - \Phi(C_{\mathrm{CLT}})) + 2C_{\mathrm{BE}}(C_{\mathrm{CLT}})\, \frac{\beta}{\sqrt{M}}\, \mathrm{TOL} \\
&= \delta + 2\frac{C_{\mathrm{BE}}(C_{\mathrm{CLT}})\, \beta}{\sigma C_{\mathrm{CLT}}}\, \mathrm{TOL}
\end{aligned}$$

This means that in the worst case, the actual error probability could be $\delta + \mathcal{O}\,(\mathrm{TOL})$ instead of $\delta$ [2]. For instance, in situations where the statistical confidence in the result is more stringent than the accuracy so that $\delta \ll \mathrm{TOL}$, the asymptotically motivated choice of $M$ in (3) could, granted the bound (4) is sharp, fail to deliver

[2]Note that $C_{\mathrm{CLT}}$ as defined in (3) grows only very slowly as $\delta$ decreases, since we have $C_{\mathrm{CLT}} < \sqrt{2\log(\delta^{-1})}$. Thus, the factor in front of TOL in the error probability can almost be neglected.

the expected level of confidence. For most r.v. however, the bound (4) is far too conservative, and one might ask whether it is reasonable to fear overshooting the error probability in the fashion we have described. The following numerical example shows the existence of r.v. for which the stopping rule (3) fails in the non-asymptotic regime

**Example 2.1.** The heavy-tailed Pareto-distribution has the probability distribution function

$$f(x) = \begin{cases} \alpha x_m^\alpha x^{-(\alpha+1)} & \text{if } x \geq x_m \\ 0 & \text{else,} \end{cases} \tag{5}$$

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

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

It is further easy to derive that for a Pareto r.v. with $\alpha = 3 + \gamma$ and $0 < \gamma < 1$,

$$\beta = \frac{\mathrm{E}\left[|X-\mu|^3\right]}{\sigma^3} = \mathcal{O}\left(\gamma^{-1}\right).$$

This implies that there exists r.v. for which the second summand of the bound (4) can become arbitrary large. So for such r.v. the stopping rule (3) might fail. Let us investigate by numerical approximations. Considering the distribution with $\alpha = 3.1$ (and $x_m = 1$), yields a heavy-tailed r.v. with known mean, variance and third moment. For a set of accuracies $\mathrm{TOL} \in [0.05, 0.2]$ and confidences $\delta = \mathrm{TOL}^\ell$, $\ell = 0.5, 1, 1.5$, and 2 we have numerically approximated $\mathrm{P}\left(\left|\overline{X}_M - \mu\right| > \mathrm{TOL}\right) \leq \delta$ using, in accordance with (3), the stopping rule

$$M = \left\lceil \frac{C_{\mathrm{CLT}}^2 \sigma^2}{\mathrm{TOL}^2} \right\rceil$$

The results, illustrated in Figure 1, show that when $\delta \ll \mathrm{TOL}$, the sought confidence is far from met.

Example 2.1 shows that for some r.v. the confidence goal of (1) will not be met by using the stopping rule (3), at least in settings with $\delta \ll \mathrm{TOL}$. Supposing we do not know the variance prior to sampling, yet another type of stopping rule failure is given in Example 2.2; it considers how the MC estimate of Algorithm 1 depends on the initial number of samples $M_0$.

**Example 2.2** (Premature Stopping)**.** In this example we will sample the mean of various r.v. using Algorithm 1 and investigate how the MC estimate $\overline{X}_M$ depends on the initial number of samples $M_0$. Let $M(M_0)$ denote the number of r.v. used in the stopped estimate as a function of the initial number of samples $M_0$. Our MC estimate goal then becomes to achieve

$$\mathrm{P}\left(\left|\overline{X}_{M(M_0)} - \mu\right| > \mathrm{TOL}\right) \leq \delta. \tag{6}$$

To investigate whether this goal is fulfilled we plot $\mathrm{P}\left(\left|\overline{X}_{M(M_0)} - \mu\right| > \mathrm{TOL}\right)$ as a function of $M_0$ for four different r.v. in Figure 2. Figure 2 indicates that the more heavy-tailed or skewed the distribution is, the higher $M_0$ is needed to ensure that the goal (6) is met.

The demonstrated stopping rule failures motivated us to study and develop ways of constructing more reliable stopping rules. In Section 3, we first analyze the stopping rule of Algorithm 1, and derive an approximate upper bound for the failure probability expressed in terms of $M, \mathrm{TOL}$ and $\delta$. In Section 4, we develop

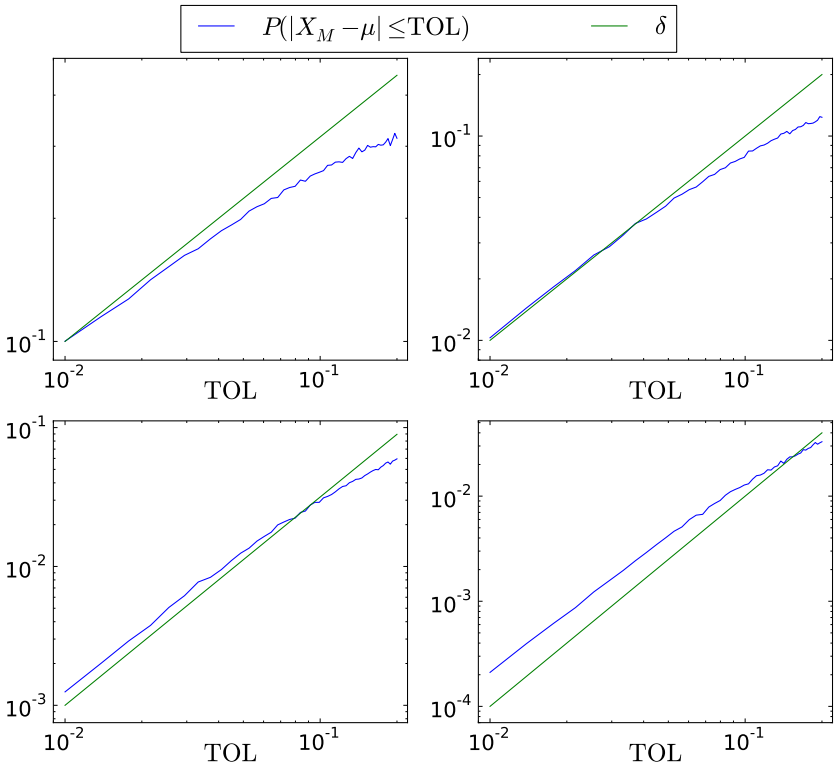


FIGURE 1. MC estimate using the stopping rule (3) for i.i.d. Pareto r.v. with parameters $\alpha = 3.1$ and $x_m = 1$. The obtained failure probability $\mathrm{P}\big(\big|\overline{X}_M - \mu\big| > \mathrm{TOL}\big)$ (blue lines) is plotted in comparison to the sought confidence parameter $\delta(\mathrm{TOL}) = \mathrm{TOL}^\ell$ (green lines), for $\ell = 0.5$ (upper left), $\ell = 1$ (upper right), $\ell = 1.5$ (lower left), and $\ell = 2$ (lower right). We observe that the smaller $\delta$ is relative to TOL, the more apparent does the failure of the stopping rule become.

a more reliable stopping rule algorithm, which in addition to second moment of the r.v. in question, also depends on third and fourth order moments. The paper is concluded with numerical examples comparing the reliability and computational cost of Algorithm 1 with the stopping rule developed in Section 4.

## 3. Error analysis for Algorithm 1

Examples 2.1 and 2.2 illustrate that for some r.v. the stopping rule in Algorithm 1 does not meet the accuracy-confidence constraint (1). To construct a more reliable stopping rule, penalty terms have to be added to the stopping criterion in Algorithm 1. Some care should be taken to make the penalty terms of right size: if too large penalties are added, the new stopping rule will be reliable but very inefficient, while if too small penalty terms are added, the algorithm will of course be efficient but unreliable.

In this section, we first derive an approximate upper bound for the failure probability

$$\mathrm{P}\Big(\big|\overline{X}_M - \mu\big| > \mathrm{TOL}\,\Big|M\Big) \tag{7}$$

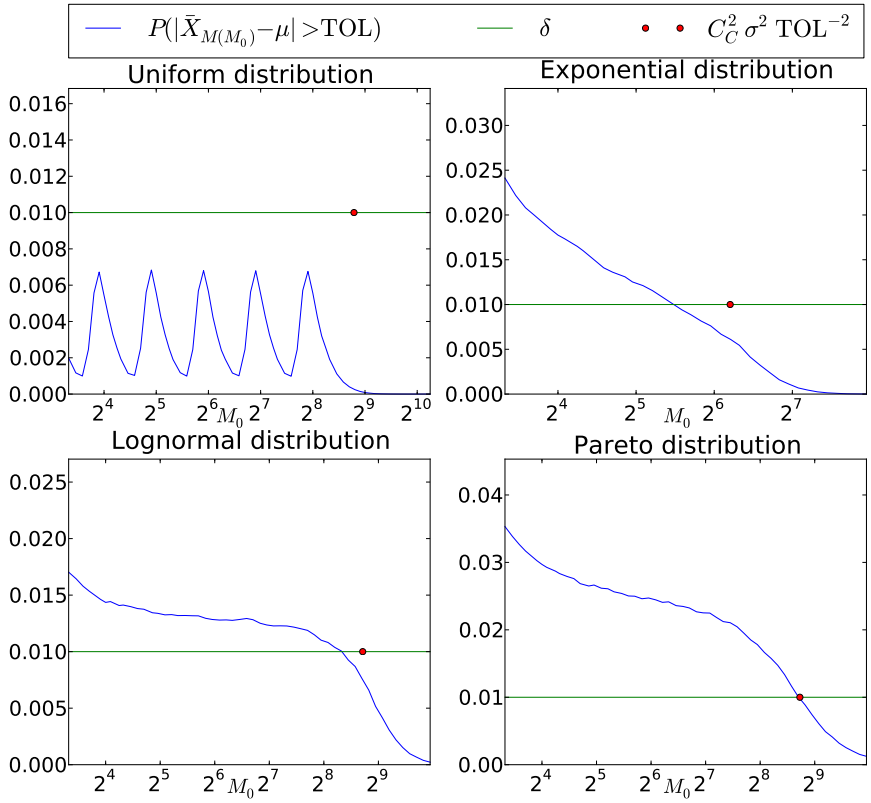


FIGURE 2. Plots of $\mathrm{P}\big(\big|\overline{X}_{M(M_0)}-\mu\big|>\mathrm{TOL}\big)$ as a function of $M_0$ when using Algorithm 1 to sample $\overline{X}_{M(M_0)}$. The accuracy and confidence is set to $\mathrm{TOL}=0.1$ and $\delta=\mathrm{TOL}^2$, respectively. **Upper left:** The Uniform distribution $X\sim U(-1,1)$ with $\mu=0$ and $\sigma^2=2/3$ (light-tailed). **Upper right:** The Exponential distribution with $\mu=1/3$ and $\sigma^2=1/9$ (not heavy-tailed). **Lower left:** The Lognormal distribution $X\sim\log(\mathcal{N}(\mu_{\mathcal{N}},\sigma^2_{\mathcal{N}}))$ with with $\mu_{\mathcal{N}}=-1$ and $\sigma^2_{\mathcal{N}}=1$. This gives $\mu=\exp(\mu_{\mathcal{N}}+\sigma^2_{\mathcal{N}}/2)$ and $\sigma^2=(\exp(\sigma^2_{\mathcal{N}})-1)\exp(2\mu_{\mathcal{N}}+\sigma^2_{\mathcal{N}})$ (quite heavy-tailed). **Lower right:** The Pareto distribution with $x_m=1$ and $\alpha=3.1$, cf. (5). (heavy-tailed).

corresponding to the stopping rule of Algorithm 1 conditional on the (random) final number of samples $M$. Clearly, the bound for (7) will also be a r.v. Using the bound for (7), we thereafter construct reasonable penalty terms to be added to the stopping criterion of our new stopping rule.

Let $\overline{\sigma}_M$ denote the sample variance generated from the stopped sample batch, i.e., the samples used to generate the output MC estimate $\overline{X}_M$. Then our first step towards an upper bound for (7) is partitioning the probability (7) into two parts

as follows

$$\begin{aligned}
&\mathrm{P}\Big(\big|\overline{X}_M - \mu\big| > \mathrm{TOL}\,\Big|M\Big) \\
&= \mathrm{P}\Big(\big|\overline{X}_M - \mu\big| > \mathrm{TOL}\,\Big|M, \big\{|\overline{\sigma}_M^2 - \sigma^2| \geq \sigma^2/2\big\}\Big)\,\mathrm{P}\Big(|\overline{\sigma}_M^2 - \sigma^2| \geq \sigma^2/2\Big|M\Big) \\
&\quad + \mathrm{P}\Big(\big|\overline{X}_M - \mu\big| > \mathrm{TOL}\,\Big|M, \big\{|\overline{\sigma}_M^2 - \sigma^2| < \sigma^2/2\big\}\Big)\,\mathrm{P}\Big(|\overline{\sigma}_M^2 - \sigma^2| < \sigma^2/2\Big|M\Big)\,.
\end{aligned} \tag{8}$$

The event $|\overline{\sigma}_M^2 - \sigma^2| \geq \sigma^2/2$ implies the estimate of the real variance is substantially wrong, and then it is likely that we use far too few samples $M$ to ensure that our MC estimate is reliable. A relatively high penalty term should therefore be added to the stopping criterion to avoid the event $|\overline{\sigma}_M^2 - \sigma^2| \geq \sigma^2/2$ from occurring. To derive such a penalty term, we will first bound the factors of the product

$$\mathrm{P}\Big(\big|\overline{X}_M - \mu\big| > \mathrm{TOL}\,\Big|M, \big\{|\overline{\sigma}_M^2 - \sigma^2| \geq \sigma^2/2\big\}\Big)\,\mathrm{P}\Big(|\overline{\sigma}_M^2 - \sigma^2| \geq \sigma^2/2\Big|M\Big) \tag{9}$$

separately. For the first factor of this product,

$$\mathrm{P}\Big(\big|\overline{X}_M - \mu\big| > \mathrm{TOL}\,\Big|M, \big\{|\overline{\sigma}_M^2 - \sigma^2| \geq \sigma^2/2\big\}\big)\Big)\,,$$

we recall that in Algorithm 1 the samples used in the output estimate $\overline{X}_M$ and for $\overline{\sigma}_M$ are independent of the samples used to determine $M$. Keeping this in mind, we derive the following approximate upper bound

$$\begin{aligned}
&\mathrm{P}\Big(\big|\overline{X}_M - \mu\big| > \mathrm{TOL}\,\Big|M, \big\{|\overline{\sigma}_M^2 - \sigma^2| \geq \sigma^2/2\big\}\Big) \\
&= \mathrm{P}\Big(\big|\overline{X}_n - \mu\big| > \mathrm{TOL}\,\Big|\big\{|\overline{\sigma}_n^2 - \sigma^2| \geq \sigma^2/2\big\}\Big)\Big|_{n=M} \\
&\lessapprox 2\left(1 - \Phi\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\sigma}\right) + C_{\mathrm{BE}}\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\sigma}\right)\frac{\beta}{\sqrt{M}}\right).
\end{aligned} \tag{10}$$

Here the Berry-Esseen bound of Corollary 1.2 was used to derive the approximate bound of the last line.

For the second factor of the product (9), we obtain the following equality

$$\mathrm{P}\Big(|\overline{\sigma}_M^2 - \sigma^2| \geq \sigma^2/2\Big|M\Big) = \mathrm{P}\big(|\overline{\sigma}_n^2 - \sigma^2| \geq \sigma^2/2\big)\big|_{n=M}\,.$$

Furthermore, using Chebycheff's inequality and k-Statistics to bound the variance of the sample variance, cf. Keeping [6], we derive that

$$\begin{aligned}
\mathrm{P}\Big(|\overline{\sigma}_M^2 - \sigma^2| \geq \sigma^2/2\Big|M\Big) &= \mathrm{P}\big(|\overline{\sigma}_n^2 - \sigma^2| \geq \sigma^2/2\big)\big|_{n=M} \\
&\leq 4\,\mathrm{E}\left[\frac{\left|\sigma^2 - \overline{\sigma}_n^2\right|^2}{\sigma^4}\right]\Bigg|_{n=M} \leq 4\frac{\sigma^4\left(\frac{2}{M-1} + \frac{\kappa}{M}\right)}{\sigma^4} = 4\left(\frac{2}{M-1} + \frac{\kappa}{M}\right).
\end{aligned}$$

Here $\kappa$ denotes the *kurtosis*, i.e.

$$\kappa = \frac{\mathrm{E}\big[|X-\mu|^4\big]}{\sigma^4} - 3.$$

We conclude that

$$\mathrm{P}\Big(|\overline{\sigma}_M^2 - \sigma^2| \geq \sigma^2/2\Big|M\Big) \leq \min\left\{1, 4\left(\frac{2}{M-1} + \frac{\kappa}{M}\right)\right\}. \tag{11}$$

Next, we want to bound the first factor of the second term of (8),

$$\mathrm{P}\Big(\big|\overline{X}_M - \mu\big| > \mathrm{TOL}\,\Big|M, \big\{|\overline{\sigma}_M^2 - \sigma^2| \leq \sigma^2/2\big\}\Big)\,.$$

The event $|\overline{\sigma}_M^2 - \sigma^2| \leq \sigma^2/2$ indicates that the variance is not substantially wrong-estimated and thereby that it is quite likely that sufficiently many samples are used in our MC estimate. Example 2.1 however illustrated that even in settings with reasonably well-estimated $M$, failing to meet the confidence is still possible. A relatively weak penalty should thus be added to the stopping criterion to avoid failure in this setting. Applying the Edgeworth expansion with truncated $o\big(n^{-1/2}\big)$ as a weak penalty, cf. Theorem 1.3, we derive the approximate bound

$$\begin{aligned}
&\mathrm{P}\Big(\big|\overline{X}_M - \mu\big| > \mathrm{TOL}\,\Big|M, \big\{|\overline{\sigma}_M^2 - \sigma^2| > \sigma^2/2\big\}\Big)\\
&= \mathrm{P}\Big(\big|\overline{X}_n - \mu\big| > \mathrm{TOL}\,\Big|\big\{|\overline{\sigma}_n^2 - \sigma^2| > \sigma^2/2\big\}\Big)\Big|_{n=M}\\
&= \mathrm{P}\left(\sqrt{n}\frac{\big|\overline{X}_n - \mu\big|}{\sigma} > \frac{\sqrt{n}\,\mathrm{TOL}}{\sigma}\,\middle|\,\big\{|\overline{\sigma}_n^2 - \sigma^2| > \sigma^2/2\big\}\right)\Bigg|_{n=M}\\
&\lessapprox 2\left(1 - \Phi\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\sigma}\right) + \frac{\left|\frac{M\,\mathrm{TOL}^2}{\sigma^2} - 1\right|\exp\left(-\frac{M\,\mathrm{TOL}^2}{\sigma^2}\right)}{6\sqrt{2\pi M}}\,\frac{\big|\mathrm{E}\big[(X-\mu)^3\big]\big|}{\sigma^3}\right).
\end{aligned} \tag{12}$$

Combining (10), (11) and (12), and noting that for all $x \in \mathbb{R}_+$ and $n \in \mathbb{N}$,

$$\frac{\big|x^2 - 1\big|\,e^{-x^2/2}}{6\sqrt{2\pi n}}\,\frac{\big|\mathrm{E}\big[(X-\mu)^3\big]\big|}{\sigma^3} \leq C_{\mathrm{BE}}(x)\,\frac{\beta}{\sqrt{n}},$$

we obtain the following approximate bound for failing to meet the accuracy of Algorithm 1 conditioned on the stopped number of samples $M$:

$$\begin{aligned}
&\mathrm{P}\Big(\big|\overline{X}_M - \mu\big| > \mathrm{TOL}\,\Big|M\Big)\\
&\lessapprox 2\left\{1 - \Phi\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\sigma}\right) + C_{\mathrm{BE}}\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\sigma}\right)\,\frac{\beta}{\sqrt{M}}\right\}\mathrm{P}\big(|\overline{\sigma}_M^2 - \sigma^2| \leq \sigma^2/2\big|\,M\big)\\
&+ 2\left\{1 - \Phi\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\sigma}\right) + \frac{\left|\frac{M\,\mathrm{TOL}^2}{\sigma^2} - 1\right|\big|\mathrm{E}\big[(X-\mu)^3\big]\big|}{\exp\left(\frac{M\,\mathrm{TOL}^2}{2\sigma^2}\right)\times 6\sqrt{2\pi M}\sigma^3}\right\}\mathrm{P}\big(|\overline{\sigma}_M^2 - \sigma^2| > \sigma^2/2\big|\,M\big)\\
&\lessapprox 2\left(1 - \Phi\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\sigma}\right)\right) + 2C_{\mathrm{BE}}\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\sigma}\right)\,\frac{\beta}{\sqrt{M}}\min\left\{1, 4\left(\frac{2}{M-1} + \frac{\kappa}{M}\right)\right\}\\
&\quad + \frac{\left|\frac{M\,\mathrm{TOL}^2}{\sigma^2} - 1\right|\big|\mathrm{E}\big[(X-\mu)^3\big]\big|}{\exp\left(\frac{M\,\mathrm{TOL}^2}{2\sigma^2}\right)\times 3\sqrt{2\pi M}\sigma^3}\max\left\{1 - 4\left(\frac{2}{M-1} + \frac{\kappa}{M}\right), 0\right\}.
\end{aligned} \tag{13}$$

## 4. A higher moments based stopping rule

From the approximate stochastic error bound (13) we will in this section construct a new, more reliable stopping rule with a stopping criterion based on second, third, and fourth moments of the r.v. that is sampled. The sampled moments our new algorithm will depend on are (here represented in biased form)

$$\begin{aligned}
\overline{\sigma}_M &:= \sqrt{\sum_{i=1}^{M}\frac{(X_i - \overline{X}_M)^2}{M}}, & \overline{\beta}_M &:= \sum_{i=1}^{M}\frac{|X_i - \overline{X}_M|^3}{M\overline{\sigma}_M^3},\\
\widehat{\beta}_M &:= \sum_{i=1}^{M}\frac{(X_i - \overline{X}_M)^3}{M\overline{\sigma}_M^3}, & \text{and}\quad \overline{\kappa}_M &:= \sum_{i=1}^{M}\frac{(X_i - \overline{X}_M)^4}{M\overline{\sigma}_M^4} - 3.
\end{aligned} \tag{14}$$

Replacing moments with sample moments in (13), we obtain a computable approximate stochastic error bound

$$\begin{aligned}
&\mathrm{P}\Big(|\overline{X}_M - \mu| > \mathrm{TOL}\,\Big|M\Big) \\
&\lessapprox 2\left(1-\Phi\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\overline{\sigma}_M}\right)\right) + 2C_{\mathrm{BE}}\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\overline{\sigma}_M}\right)\frac{\overline{\beta}_M}{\sqrt{M}}\min\left\{1, 4\left(\frac{2}{M-1}+\frac{\overline{\kappa}_M}{M}\right)\right\} \\
&\quad + \frac{\left|\frac{M\,\mathrm{TOL}^2}{\overline{\sigma}_M^2}-1\right||\widehat{\beta}_M|}{\exp\left(\frac{M\,\mathrm{TOL}^2}{2\overline{\sigma}_M^2}\right)\times 3\sqrt{2\pi M}\overline{\sigma}_M^3}\max\left\{1-4\left(\frac{2}{M-1}+\frac{\overline{\kappa}_M}{M}\right),0\right\}.
\end{aligned} \tag{15}$$

The resulting approximate stochastic error bound will be implemented as the following stopping criterion in Algorithm 2:

$$\begin{aligned}
&2\left(1-\Phi\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\overline{\sigma}_M}\right)\right) + 2C_{\mathrm{BE}}\left(\frac{\sqrt{M}\,\mathrm{TOL}}{\overline{\sigma}_M}\right)\frac{\overline{\beta}_M}{\sqrt{M}}\min\left\{1, 4\left(\frac{2}{M-1}+\frac{\overline{\kappa}_M}{M}\right)\right\} \\
&\quad + \frac{\left|\frac{M\,\mathrm{TOL}^2}{\overline{\sigma}_M^2}-1\right||\widehat{\beta}_M|}{\exp\left(\frac{M\,\mathrm{TOL}^2}{2\overline{\sigma}_M^2}\right)\times 3\sqrt{2\pi M}\overline{\sigma}_M^3}\max\left\{1-4\left(\frac{2}{M-1}+\frac{\overline{\kappa}_M}{M}\right),0\right\} < \delta
\end{aligned} \tag{16}$$

We now present the new stopping rule algorithm.

**Algorithm 2** Higher Moments Based Stopping Rule

**Input:** Accuracy TOL, confidence $\delta$, and initial number of samples $M_0$.
**Output:** $\overline{X}_M$.

Set $n = 0$, generate i.i.d. samples $\{X_i\}_{i=1}^{M_n}$ and compute the sample moments $\overline{\sigma}_{M_n}$, $\overline{\beta}_{M_n}$, $\widehat{\beta}_{M_n}$ and $\overline{\kappa}_{M_n}$ according to (14).
**while** Inequality (16) is not fulfilled. **do**
  Set $n = n + 1$ and $M_n = 2M_{n-1}$.
  Generate $M_n$ i.i.d. samples $\{X_i\}_{i=1}^{M_n}$ and compute the sample moments $\overline{\sigma}_{M_n}$, $\overline{\beta}_{M_n}$, and $\overline{\kappa}_{M_n}$.
**end while**
Set $M = M_n$, generate i.i.d. samples $\{X_i\}_{i=1}^{M}$ and return the sample mean $\overline{X}_M$.

## 5. Numerical experiments

In the numerical experiments we will estimate the mean of four differently distributed r.v. by using both the the sample variance based stopping rule in Algorithm 1 and the new higher moments based stopping rule in Algorithm 2. We compare the reliability and complexity of the algorithms, with the complexity measured in terms of average number of r.v. realizations needed to generate the given MC estimate for a given accuracy-confidence pair TOL and $\delta$. The distributions considered here are the light-tailed Uniform distribution, the Exponential distribution, the heavier-tailed Lognormal distribution, and the heavy-tailed Pareto distribution. In all these experiments we have set the algorithm parameter initial number of samples to $M_0 = 30$. From Figures 3, 4, 5, and 6, which illustrate the results of the numerical experiments, we observe that for the heavy-tailed distributions Algorithm 2 performs reliably and succeeds in meeting the accuracy-confidence constraint while

Algorithm 1 does not. For the light tailed distributions considered, both Algorithms meet the accuracy-confidence constraint. Regarding the complexity of the algorithms, we see that Algorithm 2 is only slightly more costly than Algorithm 1, and, as expected, the complexities of the algorithms seem to become more similar as TOL decreases.

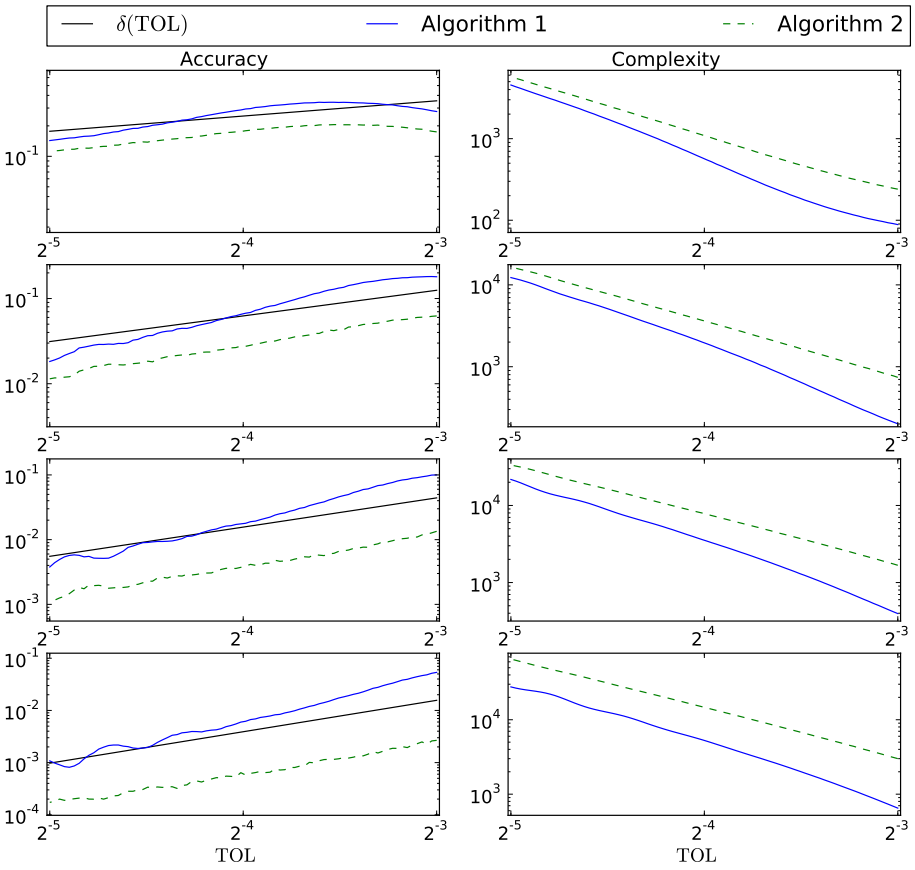


FIGURE 3. (**Pareto Distribution**) Numerical comparison of the accuracy and complexity of reaching the goal $\mathrm{P}\big(\big|\overline{X}_M - \mu\big| > \mathrm{TOL}\big) < \delta$ with Algorithm 1 and 2 when sampling Pareto distributed r.v.s with parameters $\alpha = 3.1$ and $x_m = 1$, cf. (5). Row plots from top to bottom is the output for the respective confidences $\delta(\mathrm{TOL}) = \mathrm{TOL}^{1/2}$, $\delta(\mathrm{TOL}) = \mathrm{TOL}$, $\delta(\mathrm{TOL}) = \mathrm{TOL}^{3/2}$, and $\delta(\mathrm{TOL}) = \mathrm{TOL}^2$.

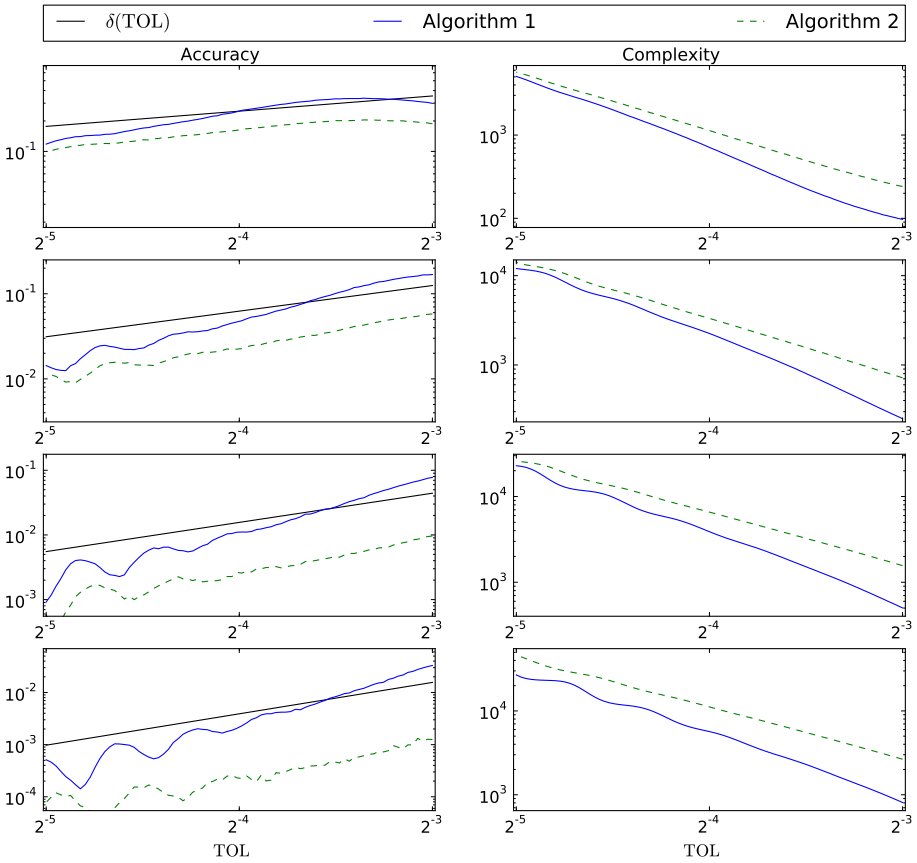


FIGURE 4. (**Lognormal Distribution**) Numerical comparison of the accuracy and complexity of reaching the goal $\mathrm{P}\big(\big|\overline{X}_M - \mu\big| > \mathrm{TOL}\big) < \delta$ with Algorithm 1 and 2 when sampling Lognormal distributed r.v. $X \sim \log(\mathcal{N}(\mu_{\mathcal{N}}, \sigma^2_{\mathcal{N}}))$ with with $\mu_{\mathcal{N}} = -1$ and $\sigma^2_{\mathcal{N}} = 1$. Row plots from top to bottom is the output for the respective confidences $\delta(\mathrm{TOL}) = \mathrm{TOL}^{1/2}$, $\delta(\mathrm{TOL}) = \mathrm{TOL}$, $\delta(\mathrm{TOL}) = \mathrm{TOL}^{3/2}$, and $\delta(\mathrm{TOL}) = \mathrm{TOL}^2$.

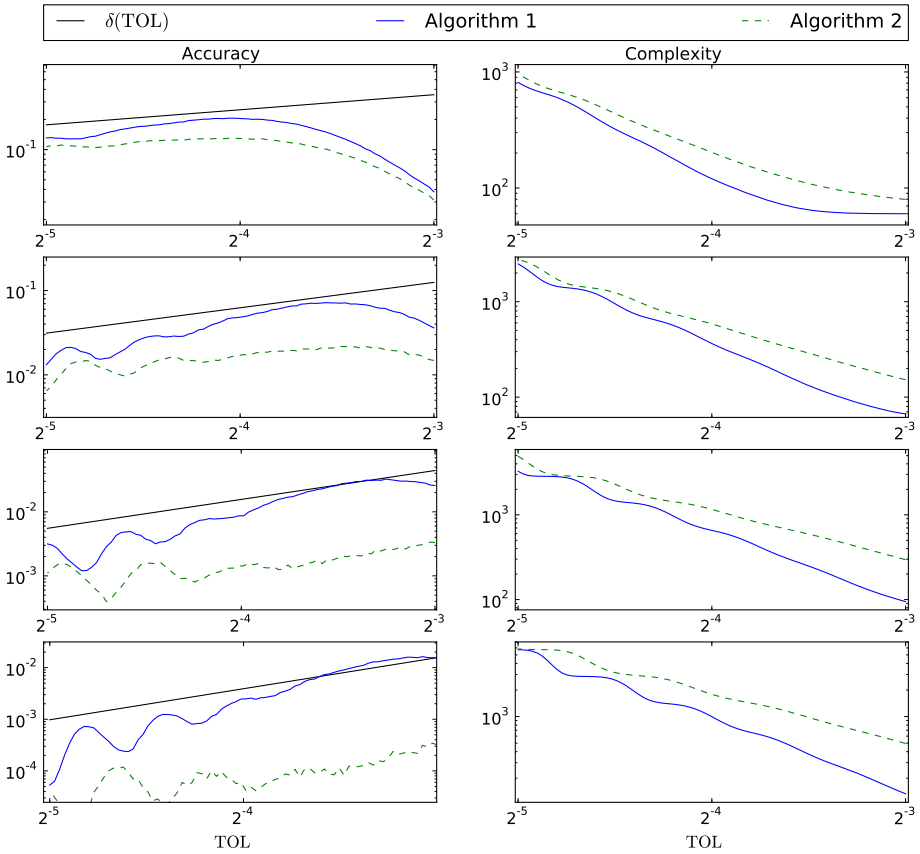


FIGURE 5. (**Exponential Distribution**) Numerical comparison of the accuracy and complexity of reaching the goal $\mathrm{P}\left(\left|\overline{X}_M - \mu\right| > \mathrm{TOL}\right) < \delta$ with Algorithm 1 and 2 when sampling exponentially distributed r.v. with $\mu = 1/3$. Row plots from top to bottom is the output for the respective confidences $\delta(\mathrm{TOL}) = \mathrm{TOL}^{1/2}$, $\delta(\mathrm{TOL}) = \mathrm{TOL}$, $\delta(\mathrm{TOL}) = \mathrm{TOL}^{3/2}$, and $\delta(\mathrm{TOL}) = \mathrm{TOL}^2$.

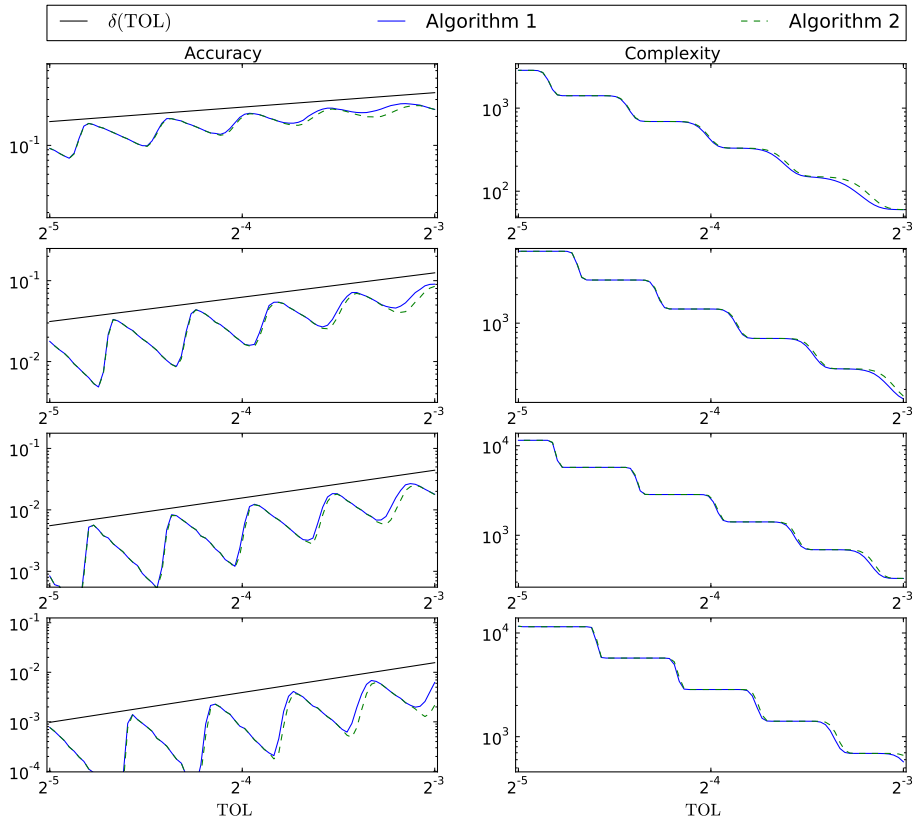


FIGURE 6. (**Uniform Distribution**) Numerical comparison of the accuracy and complexity of reaching the goal $\mathrm{P}\big(\big|\overline{X}_M - \mu\big| > \mathrm{TOL}\big) < \delta$ with Algorithm 1 and 2 when sampling uniformly distributed r.v. $X \sim U(-1,1)$. Row plots from top to bottom is the output for the respective confidences $\delta(\mathrm{TOL}) = \mathrm{TOL}^{1/2}$, $\delta(\mathrm{TOL}) = \mathrm{TOL}$, $\delta(\mathrm{TOL}) = \mathrm{TOL}^{3/2}$, and $\delta(\mathrm{TOL}) = \mathrm{TOL}^2$.

## 6. Conclusion

We have shown that second moment based sequential stopping rules such as Algorithm 1 run the risk of using too few samples in MC estimates, especially when sampling heavy-tailed r.v. in settings with very stringent confidence requirements, i.e., $\delta \ll \mathrm{TOL}$. Algorithm 2, a higher moment based stopping rule algorithm is proposed in this work, and, according to the numerical examples of Section 5, our new stopping rule performs more much reliable than Algorithm 1 while only slightly increasing the computational cost. In short, we believe that our new stopping rule presented in Algorithm 2 is well worth considering in settings with heavy tailed r.v. and/or $\delta \ll \mathrm{TOL}$.

Note that our analysis of the original Algorithm 1 critically depends on three main ingredients:

(I) a general, non-asymptotic estimate of the tail probabilities for the sample mean $\overline{X}_M$, for which we used either the non-uniform Berry-Esseen theorem given in Corollary 1.2 or the Edgeworth expansion given in Theorem 1.3,

(II) a choice between the more conservative Berry-Esseen bound and the approximate Edgeworth bound made depending on whether the sample variance of

the samples used to generate the output MC estimate is close to, or far from the true variance,

(III) an estimate of the conditional distribution function of the sample variance given the output $M$ of the stopping algorithm given in (11).

There is clearly room for improvement in all these steps. First of all, the second ingredient above is dangerous as we do not know how to estimate the correlation between $\overline{X}_M$ and the events $|\overline{\sigma}^2_M - \sigma^2| > \sigma^2/2$ and $|\overline{\sigma}^2_M - \sigma^2| \leq \sigma^2/2$. This is problematic, as these approximations can potentially have the wrong sign, i.e., it is possible that the right-hand sides of (10) and (12) are smaller than their respective left-hand sides even though we actually seek upper bounds. It is however our hope that these approximation errors are compensated by the the overly pessimistic non-uniform Berry-Esseen estimate and by using Chebycheff's inequality to bound the conditional distribution function of the sample variance. Even though the numerical evidence obtained in Section 5 seems to confirm that the compensations work well, we would prefer an analysis in which each estimation step can be controlled, at least in the sense that we indeed obtain an upper bound for the error probability.

To a lesser extent, it is not clear that the truncation of the $o\big(n^{-1/2}\big)$ of the Edgeworth expansion will lead to an upper bound for the error probability, either. In this case, the approximation error is however of higher order, so a stronger case can be made on why the effect will finally be negligible. In fact, when we used truncated Edgeworth expansion also for the estimation of (10) – instead of the non-uniform Berry-Esseen theorem – then the corresponding stopping rule turned out to be not much more reliable than Algorithm 1, indicating that there is a delicate balance between reliability in meeting the accuracy target (1) and maintaining an acceptable efficiency.

**Acknowledgments.** This work was partially supported by the Center for Predictive Computational Science), the VR project "Effektiva numeriska metoder för stokastiska differentialekvationer med tillämpningar", and King Abdullah University of Science and Technology (KAUST).

## References

[1] Paul Bratley, Bennett L. Fox, and Linus E. Schrage. *A guide to simulation (2nd ed.)*. Springer-Verlag New York, Inc., New York, NY, USA, 1987.

[2] Y. S. Chow and Herbert Robbins. On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *The Annals of Mathematical Statistics*, 36(2):pp. 457–462, 1965.

[3] W. Feller. *An introduction to probability theory and its applications. Vol II. 2nd ed.* Wiley Series in Probability and Mathematical Statistics. New York etc.: John Wiley and Sons, Inc. XXIV, 669 p. , 1971.

[4] Peter W. Glynn and Ward Whitt. The asymptotic validity of sequential stopping rules for stochastic simulations. *The Annals of Applied Probability*, 2(1):pp. 180–198, 1992.

[5] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):pp. 13–30, 1963.

[6] E. S. Keeping. *Introduction to statistical inference*. D. Van Nostrand Co., Inc., Princeton, N.J.-Toronto-London-New York, 1962.

[7] Averill M. Law and W. David Kelton. Confidence intervals for steady-state simulations, ii: A survey of sequential procedures. *Management Science*, 28(5):pp. 550–562, 1982.

[8] Averill M. Law and W. David Kelton. Confidence intervals for steady-state simulations: I. a survey of fixed sample size procedures. *Operations Research*, 32(6):pp. 1221–1239, 1984.

[9] R. Michel. On the constant in the nonuniform version of the berry-essen theorem. *Probability Theory and Related Fields*, 55:109–117, 1981. 10.1007/BF01013464.

[10] Valentin V. Petrov. *Limit theorems of probability theory*, volume 4 of *Oxford Studies in Probability*. The Clarendon Press Oxford University Press, New York, 1995. Sequences of independent random variables, Oxford Science Publications.

[11] Il'ya S Tyurin. Refinement of the upper bounds of the constants in lyapunov's theorem. *Russian Mathematical Surveys*, 65(3):586, 2010.

Department of Mathematics, University of Vienna, Nordbergstrae 15, 1090 Wien, Austria
*E-mail address*: christian.bayer@univie.ac.at

Department of Numerical Analysis, Kungl. Tekniska Högskolan, 100 44 Stockholm, Sweden
*E-mail address*: hhoel@kth.se

Division of Mathematics, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia
*E-mail address*: erik.vonschwerin@kaust.edu.sa

Division of Mathematics, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia
*E-mail address*: raul.tempone@kaust.edu.sa

# Paper IV

# HOW ACCURATE IS MOLECULAR DYNAMICS?

CHRISTIAN BAYER, HÅKON HOEL, ASHRAFUL KADIR, PETR PLECHÁČ, MATTIAS SANDBERG, ANDERS SZEPESSY, AND RAUL TEMPONE

Abstract. 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.

Contents

1. Motivation for error estimates of molecular dynamics 2
2. The Schrödinger and molecular dynamics models 3
3. A time-independent Schrödinger WKB-solution 7
3.1. Exact Schrödinger dynamics 7
3.2. Born-Oppenheimer dynamics 12
3.3. Equations for the density 12
3.4. Construction of the solution operator 13
4. Fourier integral WKB states including caustics 14
4.1. A preparatory example with the simplest caustic 14
4.2. A general Fourier integral ansatz 16
5. A global construction coupling caustics with single WKB-modes 22
5.1. Lagrangian manifolds 23
6. Computation of observables 25
7. Molecular dynamics approximation of Schrödinger observables 26
7.1. The Born-Oppenheimer approximation error 26
7.2. Why do symplectic numerical simulations of molecular dynamics work? 29
8. Analysis of the molecular dynamics approximation 29
8.1. Continuation of the construction of the solution operator 30
8.2. Stability from perturbed Hamiltonians 31
8.3. The Born-Oppenheimer approximation 34
9. Numerical examples 37
9.1. Example 1: A single WKB state 37
9.2. Example 2: A caustic state 38
10. The stationary phase expansion 43
Acknowledgment 45
References 45

2000 *Mathematics Subject Classification.* Primary: 81Q20; Secondary: 82C10.

*Key words and phrases.* Born-Oppenheimer approximation, WKB expansion, caustics, Fourier integral operators, Schrödinger operators.

The research of P.P. and A.S. was partially supported by the National Science Foundation under the grant NSF-DMS-0813893 and Swedish Research Council grant 621-2010-5647, respectively.

## 1. Motivation for error estimates of molecular dynamics

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, [14]. The wide popularity of molecular dynamics simulations relies on the fact that in many cases it agrees very well with experiments. Indeed when we have 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 quantitive accuracy estimation, without solving the Schrödinger solution. 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 computational complex to solve.

*The first step to check the accuracy* of a molecular dynamics simulation is to know what to compare with. Here we compare with the value of any *observable* $g(X)$, of nuclei positions $X$, for the *time-independent Schrödinger* eigenvalue equation $\mathcal{H}\Phi = E\Phi$, so that the approximation error we study is

$$\int_{\mathbb{R}^{3(N+n)}} g(X)\Phi(x,X)^*\Phi(x,X)\,dx\,dX - \lim_{T\to\infty}\frac{1}{T}\int_0^T g(X_t)dt\,, \tag{1.1}$$

for a molecular dynamics path $X_t$, with total energy equal to the Schrödinger eigenvalue $E$. The observable can be, for instance, the local potential energy, used in [37] to determine phase-field partial differential equations from molecular dynamics simulations, see Figure 1. The time-independent Schrödinger equation has a remarkable property of accurately predicting experiments in combination with no unknown data, thereby forming the foundation of computational chemistry. However, the drawback is the high dimensional solution space for nuclei-electron systems with several particles, restricting numerical solution to small molecules. In this paper we study the *time-independent* setting of the Schrödinger equation as the reference. The proposed approach has the advantage of avoiding the difficulty of finding the initial data for the time-dependent Schrödinger equation.

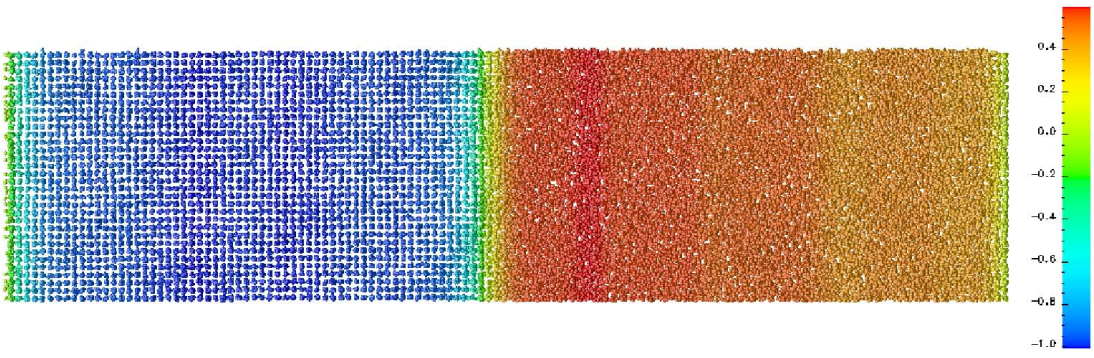


Figure 1. A Lennard-Jones molecular dynamics simulation of a phase transition with periodic boundary conditions, from [37]. The left part is solid and the right is liquid. The color measures the local potential energy.

*The second step to check the accuracy* is to derive error estimates. We have three types of error: time discretization error, sampling error and modeling error. The time discretization error comes from approximating

the differential equation for molecular dynamics with a numerical method, based on replacing time derivatives with difference quotients for a positive step size $\Delta t$. The sampling error is due to truncating the infinite $T$ and using a finite value of $T$ in the integral in (1.1). The modeling error (also called coarse-graining error) originates from eliminating the electrons in the Schrödinger nuclei-electron system and replacing the nuclei dynamics with their classical paths; this approximation error was first analyzed by Born and Oppenheimer in their seminal paper [2].

The time discretization and truncation error components are in some sense simple to handle by comparing simulations with different choice of $\Delta t$ and $T$, although it can, of course, be difficult to know that the behavior does not change with even smaller $\Delta t$ and larger $T$. The modeling error is more difficult to check since a direct approach would require to solve the Schrödinger equation. Currently the Schrödinger partial differential equation can only be solved with few particles, therefore it is not an option to solve the Schrödinger equation in general. The reason to use molecular dynamics is precisely in avoiding solution of the Schrödinger equation. Consequently the modeling error requires mathematical error analysis. In the literature there seems to be no error analysis that is precise, simple and constructive enough so that a molecular dynamics simulation can use it to asses the modeling error. Our alternative error analysis presented here is developed with the aim to to give a different point of view that could help to construct algorithms that estimate the modeling error in molecular dynamics computations. Our analysis differs from previous ones by using

- the time-independent Schrödinger equation as the reference model to compare molecular dynamics with,
- an amplitude function in a WKB-Ansatz that depends only on position coordinates $(x, X)$ (and not on momentum coordinates $(p, P)$) for caustic states,
- actual solutions of the Schrödinger equation locally and asymptotic solutions globally (and not only asymptotic solutions),
- the theory of Hamilton-Jacobi partial differential equations to derive estimates for the corresponding Hamiltonian systems, i.e., the molecular dynamics systems.

Understanding both the exact Schrödinger model and the molecular dynamics model through Hamiltonian systems allows us to obtain bounds on local and some global problems for the difference of the solutions by well-established comparison results for the solutions of Hamilton-Jacobi equations, by regarding the Schrödinger Hamiltonian and the molecular dynamics Hamiltonians as perturbations of each others. The Hamilton-Jacobi theory applied to Hamiltonian systems is inspired by the error analysis of symplectic methods for optimal control problems for partial differential equations, [31]. The result is that the modeling error can be estimated based on the difference of the Hamiltonians, for the molecular dynamics system and the Schrödinger system, along the same solution path, see Theorem 7.1 and Section 8.2. The stability analysis limits the study to non ergodic dynamics.

The next section introduces the Schrödinger and molecular dynamics models. Sections 3 and 4 establish local analysis relating the Schrödinger problem to a Hamiltonian system for non caustic and caustic states, respectively. Sections 5 and 6 extend the local picture to a global construction, in the case of non ergodic dynamics. Sections 7 and 8 formulate approximation and stability results in the Hamilton-Jacobi setting. Section 9 presents numerical results.

## 2. The Schrödinger and molecular dynamics models

In deriving the approximation of the solutions to the full Schrödinger equation the heavy particles are often treated within classical mechanics, i.e., by defining the evolution of their positions and momenta by equations of motions of classical mechanics. Therefore we denote $X_t : [0, \infty) \to \mathbb{R}^{3N}$ and $P_t : [0, \infty) \to \mathbb{R}^{3N}$ time-dependent functions of positions and momenta with time derivatives denoted by

$$\dot{X}_t = \frac{dX_t}{dt}\,, \quad \ddot{X}_t = \frac{d^2X_t}{dt^2}\,.$$

We denote the Euclidean scalar product on $\mathbb{R}^{3N}$ by

$$X \cdot Y = \sum_{i=1}^{3N} X^i Y^i\,.$$

Furthermore, we use the notation $\nabla_X \psi(x,X) = (\nabla_{X^1}\psi(x,X), \dots, \nabla_{X^N}\psi(x,X))$, and as customary $\nabla_{X^i}\psi = (\partial_{X^i_1}\psi, \partial_{X^i_2}\psi, \partial_{X^i_3}\psi)$.

On the other hand, the light particles are treated within the quantum mechanical description and the following complex valued bilinear map $\langle \cdot\,, \cdot \rangle : L^2(\mathbb{R}^{3n} \times \mathbb{R}^{3N}) \times L^2(\mathbb{R}^{3n} \times \mathbb{R}^{3N}) \to L^2(\mathbb{R}^{3N})$ will be used in the subsequent calculations

$$\langle \phi, \psi \rangle = \int_{\mathbb{R}^{3n}} \phi(x,X)^* \psi(x,X)\, dx\,. \tag{2.1}$$

The notation $\psi(x,X) = \mathcal{O}(M^{-\alpha})$ is also used for complex valued functions, meaning that $|\psi(x,X)| = \mathcal{O}(M^{-\alpha})$ holds uniformly in $x$ and $X$.

The *time-independent Schrödinger* equation

$$\mathcal{H}(x,X)\Phi(x,X) = E\Phi(x,X) \tag{2.2}$$

models many-body (nuclei-electron) quantum systems and is obtained from minimization of the energy in the solution space of wave functions, see [33, 32, 1, 35, 7]. It is an eigenvalue problem for the energy $E \in \mathbb{R}$ of the system in the solution space, described by wave functions, $\Phi : \mathbb{R}^{3n} \times \mathbb{R}^{3N} \to \mathbb{C}$, depending on electron coordinates $x = (x^1, \dots, x^n) \in \mathbb{R}^{3n}$, nuclei coordinates $X = (X^1, \dots, X^N) \in \mathbb{R}^{3N}$, and the Hamiltonian operator $\mathcal{H}(x,X)$

$$\mathcal{H}(x,X) = \mathcal{V}(x,X) - \frac{1}{2}M^{-1}\sum_{n=1}^{N} \Delta_{X^n}\,. \tag{2.3}$$

We assume that a quantum state of the system is fully described by the wave function $\Phi : \mathbb{R}^{3n} \times \mathbb{R}^{3N} \to \mathbb{C}$ which is an element of the Hilbert space of wave functions with the standard complex valued scalar product

$$\langle\!\langle \Phi, \Psi \rangle\!\rangle = \int_{\mathbb{R}^{3n}\times\mathbb{R}^{3N}} \Phi(x,X)^* \Psi(x,X)\, dx\, dX\,,$$

and the operator $\mathcal{H}$ is self-adjoint in this Hilbert space. The Hilbert space is then a subset of $L^2(\mathbb{R}^{3n} \times \mathbb{R}^{3N})$ with symmetry conditions based on the Pauli exclusion principle for electrons, see [7, 22].

In computational chemistry the operator $\mathcal{V}$, the electron Hamiltonian, is independent of $M$ and it is precisely determined by the sum of the kinetic energy of electrons and the Coulomb interaction between nuclei and electrons. We assume that the electron operator $\mathcal{V}(\cdot, X)$ is self-adjoint in the subspace with the inner product $\langle \cdot, \cdot \rangle$ of functions in (2.1) with fixed $X$ coordinate and acts as a multiplication on functions that depend only on $X$. An essential feature of the partial differential equation (2.2) is the high computational complexity of finding the solution in an antisymmetric/symmetric subset of the Sobolev space $H^1(\mathbb{R}^{3n} \times \mathbb{R}^{3N})$. The mass of the nuclei, which are much greater than one (electron mass), are the diagonal elements in the diagonal matrix $M$.

In contrast to the Schrödinger equation, a *molecular dynamics* model of $N$ nuclei $X : [0,T] \to \mathbb{R}^{3N}$, with a given potential $V_p : \mathbb{R}^{3N} \to \mathbb{R}$, can be computationally studied for large $N$ by solving the ordinary differential equations

$$\ddot{X}_t = -\nabla_X V_p(X_t)\,, \tag{2.4}$$

in the slow time scale, where the nuclei move $\mathcal{O}(1)$ in unit time. This computational and conceptual simplification motivates the study to determine the potential and its implied accuracy compared with the the Schrödinger equation, as started already in the 1920's with the Born-Oppenheimer approximation [2]. The purpose of our work is to contribute to the current understanding of such derivations by showing convergence rates under new assumptions. The precise aim in this paper is to estimate the error

$$\frac{\int_{\mathbb{R}^{3N+3n}} g(X)\Phi(x,X)^*\Phi(x,X)\, dx\, dX}{\int_{\mathbb{R}^{3N+3n}} \Phi(x,X)^*\Phi(x,X)\, dx\, dX} - \lim_{T\to\infty}\frac{1}{T}\int_0^T g(X_t)\, dt \tag{2.5}$$

for a position dependent observable $g(X)$ of the time-indepedent Schrödinger equation (2.2) approximated by the corresponding molecular dynamics observable $\lim_{T\to\infty} T^{-1}\int_0^T g(X_t)\, dt$, which is computationally cheaper to evaluate with several nuclei. The Schrödinger eigenvalue problem may typically have multiple eigenvalues and the aim is to find an eigenfunction $\Phi$ and a molecular dynamics system that can be compared.

There may be eigenfunctions that we cannot approximate and the stability analysis we use limits the study to non ergodic dynamics.

The main step to relate the Schrödinger wave function and the molecular dynamics solution is the so-called zero-order Born-Oppenheimer approximation, where $X_t$ solves the classical *ab initio* molecular dynamics (2.4) with the potential $V_p : \mathbb{R}^{3N} \to \mathbb{R}$ determined as an eigenvalue of the electron Hamiltonian $\mathcal{V}(\cdot, X)$ for a given nuclei position $X$. That is $V_p(X) = \lambda_0(X)$ and

$$\mathcal{V}(\cdot, X)\Psi_{\rm BO}(\cdot, X) = \lambda_0(X)\Psi_{\rm BO}(\cdot, X)\,,$$

for an electron eigenfunction $\Psi_{\rm BO}(\cdot, X) \in L^2(\mathbb{R}^{3n})$, for instance, the ground state. The Born-Oppenheimer expansion [2] is an approximation of the solution to the time-independent Schrödinger equation which is shown in [15, 19] to solve the time-independent Schrödinger equation approximately. This expansion, analyzed by the methods of multiple scales, pseudo-differential operators and spectral analysis in [15, 19, 13], can be used to study the approximation error (2.5). However, in the literature, e.g., [24], it is easier to find precise statements on the error for the setting of the time-dependent Schrödinger equation, since the stability issue is more subtle in the eigenvalue setting.

Instead of an asymptotic expansion we use a different method based on a Hamiltonian dynamics formulation of the *time-independent* Schrödinger eigenfunction and the stability of the corresponding perturbed Hamilton-Jacobi equations viewed as a hitting problem. This approach makes it possible to reduce the error propagation on the infinite time interval to finite time excursions from a certain co-dimension one hitting set. A motivation for our method is that it forms a sub-step in trying to estimate the approximation error using only information available in molecular dynamics simulations.

The related problem of approximating observables to the time-dependent Schrödinger equation by the Born-Oppenheimer expansions is well studied, theoretically in [4, 28] and computationally in [20] using the Egorov theorem. The Egorov theorem shows that finite time observables of the time-dependent Schrödinger equation are approximated with $\mathcal{O}(M^{-1})$ accuracy by the zero-order Born-Oppenheimer dynamics with an electron eigenvalue gap. In the special case of a position observable and no electrons (i.e., $\mathcal{V} = V(X)$ in (2.3)), the Egorov theorem states that

$$\left| \int_{\mathbb{R}^{3N}} g(X)\Phi(X,t)^*\Phi(X,t)\,dX - \int_{\mathbb{R}^{3N}} g(X_t)\Phi(X_0,0)^*\Phi(X_0,0)\,dX_0 \right| \leq C_t M^{-1}\,, \tag{2.6}$$

where $\Phi(X,t)$ is a solution to the time-dependent Schrödinger equation

$$i\partial_t\Phi(\cdot,t) = \mathcal{H}\Phi(\cdot,t)$$

with the Hamiltonian (2.3) and the path $X_t$ is the nuclei coordinates for the dynamics with the Hamiltonian $\frac{1}{2}|\dot{X}|^2 + V(X)$. If the initial wave function $\Phi(X,0)$ is the eigenfunction in (2.2) the first term in (2.6) reduces to the first term in (2.5) and the second term can also become the same in an ergodic limit. However, since we do not know that the parameter $C_t$ (bounding an integral over $(0,t)$) is bounded for all time we cannot directly conclude an estimate for (2.5) from (2.6).

In our perspective studying the time-independent instead of the time-dependent Schrödinger equation has the important differences that

- the infinite time study of the Born-Oppenheimer dynamics can be reduced to a finite time hitting problem,
- the computational and theoretical problem of specifying initial data for the Schrödinger equation is avoided, and
- computationally cheap evaluation of the position observable $g(X)$ is possible using the time average $\lim_{T\to\infty} \frac{1}{T}\int_0^T g(X_t)\,dt$ along the solution path $X_t$.

In this paper we derive the Born-Oppenheimer approximation from the time-independent Schrödinger equation (2.2) and we establish convergence rates for molecular dynamics approximations to time-independent Schrödinger observables under simple assumptions including the so-called *caustic* points, where the Jacobian determinant $\det J(X_t) \equiv \det(\partial X_t/\partial X_0)$ of the Eulerian-Lagrangian transformation of $X$-paths vanish. As mentioned previously, the main new analytical idea is an interpretation of the time-independent Schrödinger equation (2.2) as a Hamiltonian system and the subsequent analysis of the approximations by comparing Hamiltonians. This analysis employs the theory of Hamilton-Jacobi partial differential equations. The

problematic infinite-time evolution of perturbations in the dynamics is solved for non ergodic dynamics by viewing it as a finite-time hitting problem for the Hamilton-Jacobi equation, with a particular hitting set. In contrast to the traditional rigorous and formal asymptotic expansions we analyze the transport equation as a time-dependent Schrödinger equation.

The main inspiration for this paper are works [27, 6, 5] and the semi-classical WKB analysis in [25]: the works [27, 6, 5] derive the time-dependent Schrödinger dynamics of an $x$-system, $i\dot{\Psi} = \mathcal{H}_1\Psi$, from the time-independent Schrödinger equation (with the Hamiltonian $\mathcal{H}_1(x) + \epsilon\,\mathcal{H}(x, X)$) by a classical limit for the environment variable $X$, as the coupling parameter $\epsilon$ vanishes and the mass $M$ tends to infinity; in particular [27, 6, 5] show that the time derivative enters through the coupling of $\Psi$ with the classical velocity. Here we refine the use of characteristics to study classical *ab initio* molecular dynamics where the coupling does not vanish, and we establish error estimates for Born-Oppenheimer approximations of Schrödinger observables. The small scale, introduced by the perturbation

$$-(2M)^{-1}\sum_k \Delta_{X^k}$$

of the potential $\mathcal{V}$, is identified in a modified WKB eikonal equation and analyzed through the corresponding transport equation as a time-dependent Schrödinger equation along the eikonal characteristics. This modified WKB formulation reduces to the standard semi-classical approximation, see [25], in the case of the potential function $\mathcal{V} = V(X) \in \mathbb{R}$, depending only on nuclei coordinates, but becomes different in the case of operator-valued potentials studied here. The global analysis of WKB functions was initiated by Maslov in the 1960', [25], and lead to the subjects Geometry of Quantization and Quantum Ergodicity, relating global classical paths to eigenfunctions of the Schrödinger equation, see [10] and [38]. The analysis presented in this paper is based on a Hamiltonian system interpretation of the time-independent Schrödinger equation. Stability of the corresponding Hamilton-Jacobi equation, bypasses the usual separation of nuclei and electron wave functions in the time-dependent self-consistent field equations, [3, 23, 36].

A unique property of the time-independent Schrödinger equation we use is the interpretation that the dynamics $X_t \in \mathbb{R}^{3N}$ can return to a co-dimension one surface $I$ which then can reduce the dynamics to a hitting time problem with finite-time excursions from $I$. We assume that the (Lagrangian) manifold, generated by the visited points $(X_t, \dot{X}_t) \in \mathbb{R}^{6N}$ in phase space is smooth, which excludes ergodic dynamics. Another advantage of viewing the molecular dynamics as an approximation of the eigenvalue problem is that stochastic perturbations of the electron ground state can be interpreted as a Gibbs distribution of degenerate nuclei-electron eigenstates of the Schrödinger eigenvalue problem (2.2), see [34]. The time-independent eigenvalue setting also avoids the issue on "wave function collapse" to an eigenstate, present in the time-dependent Schrödinger equation.

Theorem 7.1 demonstrates that observables from the zero-order Born-Oppenheimer dynamics approximate observables for the Schrödinger eigenvalue problem with the error of order $\mathcal{O}(M^{-1+\delta})$, for any $\delta > 0$, assuming that the electron eigenvalues satisfy a spectral gap condition and that the Lagrangian manifold is smooth. The result is based on the Hamiltonian (2.3) with any potential $\mathcal{V}$ that is smooth in $X$, e.g., a regularized version of the Coulomb potential. The derivation does not assume that the nuclei are supported on small domains; in contrast derivations based on the time-dependent self-consistent field equations require nuclei to be supported on small domains. The reason that the small support is not needed here comes from the combination of the characteristics and sampling from an equilibrium density. In other words, the nuclei paths behave classically although they may not be supported on small domains. Section 6 shows that caustics couple the WKB modes, as is well-known from geometric optics, see [18, 25], and generate non-orthogonal WKB modes that are coupled in the Schrödinger density. On the other hand, with a spectral gap and without caustics the Schrödinger density is asymptotically decoupled into a simple sum of individual WKB densities. Section 4 constructs a WKB-Fourier integral Schrödinger solution for caustic states. Section 7.2 relates the approximation results to the accuracy of symplectic numerical methods for molecular dynamics.

We believe that these ideas can be further developed to better understanding of molecular dynamics simulations. Our study does not directly apply to ergodic dynamics, since then the Lagrangian manifold becomes dense in a set of dimension $6N - 1$ in phase-space, which violates our assumption of a smooth Lagrangian manifold of dimension $3N$. It would also be desirable to have more precise conditions on the data (i.e. molecular dynamics initial data and potential $\mathcal{V}$) instead of our implicit assumption on hitting

times, smooth Lagrangian manifold and convergence of the Born-Oppenheimer power series approximation in Lemma 8.2.

## 3. A time-independent Schrödinger WKB-solution

3.1. **Exact Schrödinger dynamics.** For the sake of simplicity we assume that all nuclei have the same mass. If this is not the case, we can introduce new coordinates $M_1^{1/2}\tilde{X}^k = M_k^{1/2}X^k$, which transform the Hamiltonian to the form we want $\mathcal{V}(x, M_1^{1/2}M^{-1/2}\tilde{X}) - (2M_1)^{-1}\sum_{k=1}^N \Delta_{\tilde{X}^k}$. The singular perturbation $-(2M)^{-1}\sum_k \Delta_{X^k}$ of the potential $\mathcal{V}$ introduces an additional small scale $M^{-1/2}$ of high frequency oscillations, as shown by a WKB-expansion, see [29, 17, 16, 26]. We shall construct solutions to (2.2) in such a WKB-form

$$\Phi(x, X) = \phi(x, X)e^{iM^{1/2}\theta(X)}\,, \tag{3.1}$$

where the amplitude function $\phi : \mathbb{R}^{3n} \times \mathbb{R}^{3N} \to \mathbb{C}$ is complex valued, the phase $\theta : \mathbb{R}^{3N} \to \mathbb{R}$ is real valued, and the factor $M^{1/2}$ is introduced in order to have well-defined limits of $\phi$ and $\theta$ as $M \to \infty$. Note that it is trivially always possible to find funtions $\phi$ and $\theta$ satisfying (3.1), even in the sense of a true equality. Of course, the ansatz only makes sense if $\phi$ and $\theta$ do not have strong oscillations for large $M$. The standard WKB-construction, [25, 16], is based on a series expansion in powers of $M^{1/2}$ which solves the Schrödinger equation with arbitrary high accuracy. Instead of an asymptotic solution, we introduce an actual solution based on a time-dependent Schrödinger transport equation. This transport equation reduces to the formulation in [25] for the case of a potential function $\mathcal{V} = V(X) \in \mathbb{R}$, depending only on nuclei coordinates $X \in \mathbb{R}^{3N}$, and modifies it for the case of a self-adjoint potential operator $\mathcal{V}(\cdot, X)$ on the electron space $L^2(\mathbb{R}^{3n})$ which is the primary focus of our work here. In Sections 6 and 4 we use a linear combination of WKB-eigensolutions, but first we study the simplest case of a single WKB-eigensolution as motivated by the following subsection.

3.1.1. *Molecular dynamics from a piecewise constant electron operator on a simplex mesh.* The purpose of this section is to convey a first formal understanding of the relation between ab initio molecular dynamics $\ddot{X}_t = -\nabla_X\lambda_0(X_t)$ and the Schrödinger eigenvalue problem (2.2) and motivate the WKB ansatz (3.1). In subsequent sections we will describe precise analysis of error estimates for the WKB-method. The idea behind this first study is to approximate the electron operator $\mathcal{V}$ by a finite dimensional matrix $\mathcal{V}^h$, which is piecewise constant on a simplex mesh in the variable $X$, with the mesh size $h$. Furthermore, we introduce the change of variables

$$\Phi = \sum_{j=0}^{J} \varphi_j\Psi_j =: \Psi\varphi$$

based on the piecewise constant electron eigenvalues and eigenvectors $\mathcal{V}^h\Psi_j = \lambda_j^h\Psi_j$, $\langle\Psi_j, \Psi_j\rangle = 1$, $j = 0, \ldots J$, normalized and ordered with respect to increasing eigenvalues. Then the Schrödinger equation (2.2) becomes

$$-\frac{1}{2M}\Delta_X(\Psi\varphi) + \mathcal{V}^h\Psi\varphi = E\Psi\varphi\,,$$

with the notation $\Delta_X = \sum_j \Delta_{X_j}$, so that on each simplex

$$-\frac{1}{2M}\Delta_X\varphi_j + \lambda_j^h\varphi_j = E\varphi_j\,,$$

which by separation of variables, for each $j = 0, 1, 2, \ldots, J$, implies

$$\varphi_j = \sum_{P^j} a(P^j)e^{iM^{1/2}P^j\cdot X} \tag{3.2}$$

for any $P^j \in \mathbb{C}^{3N}$ that satisfies the eikonal equation

$$\frac{1}{2}P^j \cdot P^j + \lambda_j^h = E\,,$$

for any $a(P^j) \in \mathbb{C}$, if all components of $P^j$ are non zero. If $P_k^j = 0$ we have $a(P^j) = \prod_{\{k\,:\,P_k^j=0\}}(A_kX_k + B_k)$ for any $A_k \in \mathbb{C}, B_k \in \mathbb{C}$, since $e^{\pm iM^{1/2}P_k^jX_k} = 1$ in this case. The solution $\Phi$, to (2.2), and its normal

derivative are continuous at the interfaces of the simplices. On the intersection of the faces the normal derivative is not defined but this set is of measure zero and thus negligible as seen from the $H^1(\mathbb{R}^{3N})$ solution concept of (2.2).

We investigate a simpler, one-dimensional case, $X \in \mathbb{R}$, first. Then the solution $\varphi$ simplifies to

$$\varphi_j = a_j e^{iM^{1/2}P^j \cdot X} + b_j e^{-iM^{1/2}P^j \cdot X}$$

for $a_j, b_j, P^j \in \mathbb{C}$ and $(P^j)^2/2 + \lambda_j = E$. The continuity conditions

$$\begin{aligned} \lim_{X \to X_0+} \Phi(X) &= \lim_{X \to X_0-} \Phi(X) \\ \lim_{X \to X_0+} \partial_X \Phi(X) &= \lim_{X \to X_0-} \partial_X \Phi(X) \end{aligned} \tag{3.3}$$

hold for any $X_0 \in \mathbb{R}$, in particular, at the interval boundary where for $X_0 = 0$

$$\begin{aligned} \lim_{X \to X_0\pm} \Phi(X) &= \sum_j (a_{j\pm}\Psi_{j\pm} + b_{j\pm}\Psi_{j\pm}) \\ \lim_{X \to X_0\pm} \partial_X \Phi(X) &= iM^{1/2} \sum_j (a_{j\pm}P^j_\pm \Psi_{j\pm} - b_{j\pm}P^j_\pm \Psi_{j\pm})\,. \end{aligned} \tag{3.4}$$

It is clear that given $a_-$ and $b_-$ we can determine $a_+$ and $b_+$ so that (3.3) holds. In order to prepare for the multi-dimensional case it is convenient to consider each incoming wave $a_-$ and $b_+$ separately: the incoming $a_-$ wave is split into a refracted $a_+$ and reflected $b_-$ wave

$$\sum_j a_{j-}\Psi_{j-}P^j_- = \sum_j (a_{j+}\Psi_{j+}P^j_+ + b_{j-}\Psi_{j-}P^j_-) \tag{3.5}$$

and similarly the incoming $b_+$ wave is split into a refracted $b_-$ wave and a reflected $a_+$ wave, see Figure 2. The jump conditions at the different interfaces are coupled by the oscillatory functions $e^{\pm iM^{1/2}P^j \cdot X}$. The global construction of $\varphi$ and $\Psi$ in one dimension follows by marching in the positive $X$-direction to successive intervals, creating in each interval both a $e^{iM^{1/2}P^j \cdot X}\Psi_j$ and a $e^{-iM^{1/2}P^j \cdot X}\Psi_j$ wave.

In general each interface condition (3.4) also couples all eigenvectors $\Psi_j$. However, we shall see that if $M$ is large, $\mathcal{V}$ smooth and there is a spectral gap $\lambda_1 - \lambda_0 > c > 0$ then, in the limit of the simplex size $h$ tending to zero, there is an asymptotically uncoupled WKB-solution $\Phi(x, X) = \phi(x, X)e^{iM^{1/2}\theta(X)}$, where $\theta : \mathbb{R}^{3N} \to \mathbb{R}$, $\phi : \mathbb{R}^{3n} \times \mathbb{R}^{3N} \to \mathbb{C}$. Under these assumptions the Born-Oppenheimer approximation in Lemma 8.2 shows that $\phi$ is asymptotically parallel, in $L^2(dx)$, to the electron eigenfunction $\Psi_0$ as $M \to \infty$. The gradient $\nabla_X \theta(X) = P^0$ is obtained from the differential $\theta(X) = \theta(X_0) + \nabla_X \theta(X_0) \cdot (X - X_0) + o(|X - X_0|)$.

In the case of electron eigenvalue crossing, i.e., $\lambda_1(X) = \lambda_0(X)$ for some $X$, or so called avoided crossings (meaning that the eigenvalue gap $c \ll 1$ is small and dependent on $M$), a refraction will, in general, include all components $a_j e^{iM^{1/2}P^j \cdot X}\Psi_j$, $j = 1, \ldots, J$ and consequently the Born-Oppenheimer approximation fails.

The construction of a solution to the Schrödinger equation with a piecewise constant potential is more involved in the multi-dimensional case for two reasons: each reflection at an interface generates, in general, an additional path in a new direction, so that many paths are needed. Furthermore, the construction of a solution to the eikonal equation is more complicated since the jump condition (3.4) implies that the tangential component $P^j_t$ of $P^j$ must be continuous across a simplex face and only the normal component $P^j_n = P^j - P^j_t$ may have a jump. In multi-dimensional cases it is still possible to construct a solution of the form (3.2) by following the characteristic paths $\dot{X}_t = P^j(X_t)$ and using the jump conditions (3.4): when the path $X_t$ hits a simplex face, the tangential part $P^j_t$ of $P^j$ is continuous and the normal component $P^j_n$ of $P^j$ may jump. At a simplex face the new value of the $P^j_n$ is determined by $(P^j_n \cdot P^j_n + P^j_t \cdot P^j_t)/2 + \lambda^h_j = E$. Analogously to the one dimensional case we treat the pair $e^{iM^{1/2}(P^j_t + P^j_n)\cdot X}$ and $e^{iM^{1/2}(P^j_t - P^j_n)\cdot X}$ together. However, each collision with $e^{iM^{1/2}(P^j_t + P^j_n)\cdot X}$ on an interface now creates a reflected wave in another direction, in particular, $e^{iM^{1/2}(P^j_t - P^j_n)\cdot X}\Psi_j$, and we get many paths to follow. Therefore each mode $e^{iM^{1/2}P^j \cdot X}$ follows its characteristic $X_t$, where $\dot{X}_t = P^j$, through the simplex to the adjacent simplicial faces, which the characteristic pass through when they leave the simplex, and at these outflow faces a reflected mode is created and a refracted mode continues into the adjacent simplices, see Figure 2. In this way we can

formally construct a solution of the form $\sum_{P^j} a(P^j)e^{iM^{1/2}P^j\cdot X}\Psi_j$ to the Schrödinger equation (2.2), with possibly several different characteristic paths in each simplex.

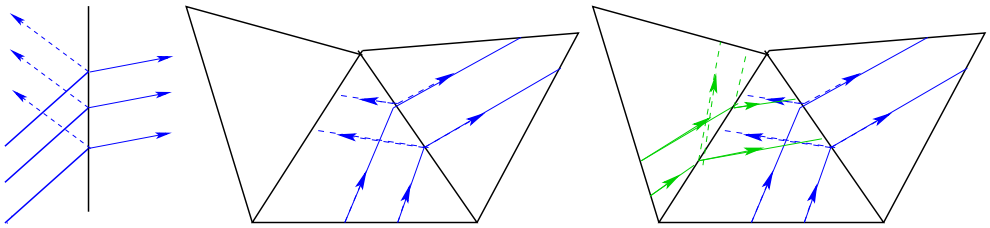

FIGURE 2. The value of $P^j$ is constructed by following the characteristic paths $X_t$ (the blue and green curves), based on $\dot{X}_t = P^j$, with a reflection-refraction at each simplex face (left) following the path through simplices (middle) and each simplex may have several $P^j$ (right).

In conclusion, the piecewise constant electron operator shows that the solution to the Schrödinger equation (2.2) is composed of a linear combination of highly oscillatory function modes $a_j e^{iM^{1/2}P^j\cdot X}\Psi_j$ based on the electron eigenvectors $\Psi_j$ and eigenvalues $\lambda_j$, where $P^j$ satisfies the eikonal equation $P^j\cdot P^j/2+\lambda_j(X) = E$. These modes can be followed be characteristics $\dot{X} = P^j$ from simplex to simplex. In this paper we show that observables based on the related WKB Schrödinger solutions can be approximated by molecular dynamics time averages, when there is a spectral gap around $\lambda_0$.

3.1.2. *A first WKB-solution.* The WKB-solution satisfies the Schrödinger equation (2.2) provided that

$$
\begin{aligned}
0 &= (\mathcal{H}-E)\phi\; e^{iM^{1/2}\theta(X)} \\
&= \left((\frac{1}{2}|\nabla_X\theta|^2+\mathcal{V}-E)\phi - \frac{1}{2M}\Delta_X\phi - \frac{i}{M^{1/2}}(\nabla_X\phi\cdot\nabla_X\theta+\frac{1}{2}\phi\,\Delta_X\theta)\right) e^{iM^{1/2}\theta(X)}\,.
\end{aligned}
\tag{3.6}
$$

We shall see that only eigensolutions $\Phi$ that correspond to dynamics without caustics correspond to such a single WKB-mode, as for instance when the eigenvalue $E$ is inside an electron eigenvalue gap. Solutions in the presence of caustics use a Fourier integral of such WKB-modes, and we treat this case in detail in Section 4. To understand the behavior of $\theta$, we multiply (3.6) by $\phi^* e^{-iM^{1/2}\theta(X)}$ and integrate over $\mathbb{R}^{3n}$. Similarly we take the complex conjugate of (3.6), and multiply by $\phi\, e^{iM^{1/2}\theta(X)}$ and integrate over $\mathbb{R}^{3n}$. By adding these two expressions we obtain

$$
\begin{aligned}
0 = 2\big(\frac{1}{2}|\nabla_X\theta|^2 - E\big)\,\langle\phi,\phi\rangle + \underbrace{\langle\phi,\mathcal{V}\phi\rangle + \langle\mathcal{V}\phi,\phi\rangle}_{=2\langle\phi,\mathcal{V}\phi\rangle} - \frac{1}{2M}\left(\langle\phi,\Delta_X\phi\rangle + \langle\Delta_X\phi,\phi\rangle\right) \\
- \frac{i}{M^{1/2}}\underbrace{\left(\langle\phi,\nabla_X\phi\cdot\nabla_X\theta\rangle - \langle\nabla_X\phi\cdot\nabla_X\theta,\phi\rangle\right)}_{=2i\mathrm{Im}\,\langle\phi,\nabla_X\phi\cdot\nabla_X\theta\rangle} + \frac{i}{2M^{1/2}}\underbrace{\big(\langle\phi,\phi\rangle - \langle\phi,\phi\rangle\big)}_{=0}\Delta_X\theta\,.
\end{aligned}
\tag{3.7}
$$

The purpose of the phase function $\theta$ is to generate an accurate approximation in the limit as $M\to\infty$. A possible and natural definition of $\theta$ would be the formal limit of (3.7) as $M\to\infty$, which is the *Hamilton-Jacobi equation*, also called the *eikonal equation*

$$
\frac{1}{2}|\nabla_X\theta|^2 = E - V_0\,,
\tag{3.8}
$$

where the function $V_0 : \mathbb{R}^{3N}\to\mathbb{R}$ is

$$
V_0 := \frac{\langle\phi,\mathcal{V}\phi\rangle}{\langle\phi,\phi\rangle}\,.
\tag{3.9}
$$

The solution to the Hamilton-Jacobi eikonal equation can be constructed from the associated Hamiltonian system

$$
\begin{aligned}
\dot{X}_t &= P_t \\
\dot{P}_t &= -\nabla_X V_0(X_t)
\end{aligned}
\tag{3.10}
$$

through the characteristics path $(X_t, P_t)$ satisfying $\nabla_X\theta(X_t) =: P_t$. The amplitude function $\phi$ can be determined by requiring the ansatz (3.6) to be a solution, which gives

$$\begin{aligned} 0 &= (\mathcal{H}-E)\phi\, e^{iM^{1/2}\theta(X)} \\ &= \Big( \underbrace{(\frac{1}{2}|\nabla_X\theta|^2 + V_0 - E)}_{=0}\,\phi \\ &\quad -\frac{1}{2M}\Delta_X\phi + (\mathcal{V}-V_0)\phi - \frac{i}{M^{1/2}}(\nabla_X\phi\cdot\nabla_X\theta + \frac{1}{2}\phi\Delta_X\theta)\Big)\, e^{iM^{1/2}\theta(X)}\,, \end{aligned}$$

so that by using (3.8) we have

$$-\frac{1}{2M}\Delta_X\phi + (\mathcal{V}-V_0)\phi - \frac{i}{M^{1/2}}(\nabla_X\phi\cdot\nabla_X\theta + \frac{1}{2}\phi\Delta_X\theta) = 0\,.$$

The usual method for determining $\phi$ from this so-called *transport equation* uses an asymptotic expansion $\phi \simeq \sum_{k=0}^{K} M^{-k/2}\phi_k$, see [15, 19] and the beginning of Section 8. An alternative is to write it as a Schrödinger equation, similar to work in [25]: we apply the characteristics in (3.10) to write

$$\frac{d}{dt}\phi(X_t) = \nabla_X\phi\cdot\dot{X}_t = \nabla_X\phi\cdot\nabla_X\theta\,,$$

and define the weight function $G$ by

$$\frac{d}{dt}\log G_t = \frac{1}{2}\Delta_X\theta(X_t)\,, \tag{3.11}$$

and the variable $\psi_t := \phi(X_t)G_t$. We use the notation $\phi(X)$ instead of the more precise $\phi(\cdot, X)$, so that e.g. $\psi_t = \psi_t(x) = \phi(x, X_t)G_t$. Then the transport equation becomes a Schrödinger equation

$$iM^{-1/2}\dot{\psi}_t = (\mathcal{V}-V_0)\psi_t - \frac{G_t}{2M}\Delta_X\left(\frac{\psi_t}{G_t}\right)\,. \tag{3.12}$$

In conclusion, equations (3.8)-(3.12) determine the WKB-ansatz (3.1) to be a local solution to the Schrödinger equation (2.2).

**Theorem 3.1.** *Assume the Hamilton-Jacobi equation, with the corresponding Hamiltonian,*

$$H_{\rm S}(X,P) := \frac{1}{2}|P|^2 + \underbrace{\frac{\langle\psi(X), \mathcal{V}(X)\psi(X)\rangle}{\langle\psi(X),\psi(X)\rangle}}_{=:V_0(X)} - E = 0\,,$$

*based on the primal variable $X$ and the dual variable $P = P(X) = \nabla_X\theta(X)$, has a smooth solution $\theta(X)$ in a domain $\mathcal{U} \subseteq \mathbb{R}^{3N}$, then $\theta$ generates a solution to the time-independent Schrödinger equation $(\mathcal{H}-E)\Phi = 0$ in $\mathcal{U}$, in the sense that*

$$\Phi(X_t, x) = \hat{G}^{-1}(X_t)\hat{\psi}(x, X_t)e^{iM^{1/2}\theta(X_t)}\,,$$

*solves the equation* (2.2) *in $\mathcal{U}$, where $\hat{\psi}(X_t) := \psi_t$ satisfies the transport equation* (3.12) *and*

$$\begin{aligned} &\hat{G}(X_t) = G_t\,, \\ &\frac{d}{dt}\log G_t = \frac{1}{2}\Delta_X\theta(X_t)\,, \\ &(X_t, P_t) \text{ solves the Hamiltonian system } (3.10) \text{ corresponding to } H_{\rm S}. \end{aligned}$$

The theorem tells us that if there is a $C^2$ solution to the Hamilton-Jacobie equation, then we have a family of characteristic paths with the desired property. It is well know that Hamilton-Jacobi equations in general do not have global $C^2$ solutions, due to $X$-paths that collide, as seen by (5.4) generating blow up in $\partial_{XX}\theta(X)$. However if the domain is small enough, the data on the boundary is compatible (in the sense that $H_S(X, \nabla_X\theta(X)) = 0$ in the boundary) and noncharacteristic (in the sense that the normal derivative $\partial_n\theta(X) \neq 0$ on the boundary) and $V_0$ is smooth, then the converse property holds: that the characteristics generate a local solution to the Hamilton-Jacobi equation, see Ref. [12]. In Section 5 we describe Maslov's method to find a global asymptotic solution by patching together local solutions; an important ingredient is how to set up data for the Hamiltonian system, which is previewed in the next section.

3.1.3. *Data for the Hamiltonian system.* For the energy $E$ chosen larger than the potential energy, that is such that $E \geq V_0$, Theorem 3.1 yields a solution $\theta : \mathcal{U} \to \mathbb{R}$ to the eikonal equation (3.8) locally in a neighborhood $\mathcal{U} \subseteq \mathbb{R}^{3N}$, for regular compatible data $(X_0, P_0)$ given on a $3N-1$ dimensional "inflow"-domain $X_0 \in I \subset \overline{\mathcal{U}}$. For a Schrödinger eigenvalue problem, the domain $I$ and the data $(X_0, P_0)|_I$ are not given (except that the total energy is $E$). In contrast for a scattering problem, the domain $I$ has given data. If paths leaving from $I$ return to $I$, there is an additional global compatibility of data on $I$: assume $X_0 \in I$ and $X_t \in I$, then the values $P_t$ are determined from $P_0$; continuing the path to subsequent hitting points $X_{t_j} \in I$, $j = 1, 2, \ldots$ determines $P_{t_j}$ from $P_0$. The characteristic path $(X_t, P_t)$, $t > 0$, generates a $3N$ dimensional Lagrangian manifold in the $6N$ dimensional phase space $(X, P)$, which is smooth under our assumptions. This Lagrangian manifold is in general only locally of the form $(X, P(X))$, but in the case of no caustics it is globally of this form and then there is a phase function $X \mapsto \theta(X)$ such that $P(X) = \nabla_X \theta(X)$ globally. Section 5.1 reviews background material on Lagrangian manifolds used in this paper.

In Section 4 we study phase space manifolds with caustics and Section 5 presents a global construction of a Lagrangian manifold in some cases. We will use a variant of Maslov's construction [25] to obtain a global asymptotic solution to the Schrödinger equation (2.2) from local WKB-solutions and we apply a Poincare map to determine the initial Lagrangian manifold, as described in Section 5: the first step is to define a codimension one hitting plane in the phase space $\mathbb{R}^{6N}$; the problem is reduced to find an initial Lagrangian manifold of dimension $3N-1$ in the hitting plane by following the characteristic paths extending $\theta_j$ and $\phi_j$ locally; around a caustic the solution is a Fourier integral of WKB solutions, described in Section 4 and the stationary phase method yields boundary conditions for the phase and amplitude functions from the Fourier integral solution; on the hitting plane the solution has to coincide with the initial data, giving a fixed point problem for the initial Lagrangian manifold.

3.1.4. *Liouville's formula.* In this section we verify Liouville's formula

$$\frac{G_0^2}{G_t^2} = e^{-\int_0^t \operatorname{Tr}\left(\nabla_X P(X_t)\right) dt} = \left| \det \frac{\partial(X_0)}{\partial(X_t)} \right| , \tag{3.13}$$

given in [25]. The characteristic $\dot{X}_t = P(X_t)$ implies $\frac{d}{dt} J(X_t) = \nabla_X P \, J(X)$, where $J(X)_{ij} = \partial X_t^i / \partial X_0^j$ denotes the first variation with respect to perturbations of the initial data. The logarithmic derivative then satisfies $d/dt\big(\log J(X)\big)_{ij} = \partial_{X^j} P^i(X_t) = \partial_{X^i X^j} \theta(X)$ which implies that $\log J(X_t)$ is symmetric and shows that (3.13) holds

$$\operatorname{div} P = \operatorname{Tr} \nabla_X P = \frac{d}{dt} \operatorname{Tr} \log J(X) = \frac{d}{dt} \log \det J(X) \, .$$

The last step uses that $J(X)$ can be diagonalized by an orthogonal transformation and that the trace is invariant under orthogonal transformations.

3.1.5. *The density and the first variation.* Note that the nuclei density, using $\hat{G}$, can be written

$$\rho := \frac{\langle \phi, \phi \rangle}{\int_{\mathbb{R}^{3N}} \langle \phi, \phi \rangle \, dX} = \frac{\langle \hat{\psi}, \hat{\psi} \rangle \, \hat{G}^{-2}}{\int_{\mathbb{R}^{3N}} \langle \hat{\psi}, \hat{\psi} \rangle \, \hat{G}^{-2} \, dX} \, , \tag{3.14}$$

and since each time $t$ determines a unique point $(X_t, P_t) = (X_t, \nabla_X \theta(X_t))$ in the phase space the functions $\hat{G}$ and $\hat{\psi}$ are well defined.

The integrating factor $G$ and its derivative $\partial_{X^i} G$ can be determined from $(P, \partial_{X^i} P, \partial_{X^i X^j} P)$ along the characteristics by the following characteristic equations obtained from (3.8) by differentiation with respect

to $X$

$$\begin{aligned}\frac{d}{dt}\partial_{X^r}P^k &= \left[\sum_j P^j\partial_{X^jX^r}P^k = \sum_j P^j\partial_{X^rX^k}P^j\right]\\ &= -\sum_j \partial_{X^r}P^j\partial_{X^k}P^j - \partial_{X^rX^k}V_0\,,\\ \frac{d}{dt}\partial_{X^rX^q}P^k &= \left[\sum_j P^j\partial_{X^jX^rX^q}P^k + \sum_j P^j\partial_{X^rX^kX^q}P^j\right]\\ &= -\sum_j \partial_{X^r}P^j\partial_{X^kX^q}P^j - \sum_j \partial_{X^rX^q}P^j\partial_{X^k}P^j - \partial_{X^rX^kX^q}V_0\,,\end{aligned} \tag{3.15}$$

and similarly $\partial_{X^iX^j}G$ can be determined from $(P, \partial_{X^i}P, \partial_{X^iX^j}P, \partial_{X^iX^jX^k}P)$.

3.2. **Born-Oppenheimer dynamics.** The Born-Oppenheimer approximation leads to the standard formulation of *ab initio* molecular dynamics, in the micro-canonical ensemble with the constant number of particles, volume and energy, for the nuclei positions $X = X_{\rm BO}$,

$$\begin{aligned}\dot X_t &= P_t\,,\\ \dot P_t &= -\nabla_X\lambda_0(X_t)\,,\end{aligned} \tag{3.16}$$

by using that the electrons are in the eigenstate $\psi = \Psi_{\rm BO}$ with eigenvalue $\lambda_0$ to $\mathcal{V}$, in $L^2(dx)$ for fixed $X$, i.e., $\mathcal{V}(X)\Psi_{\rm BO} = \lambda_0(X)\Psi_{\rm BO}$. The corresponding Hamiltonian is $H_{\rm BO}(X,P) := |P|^2/2 + \lambda_0(X)$ with the eikonal equation

$$\frac{1}{2}|\nabla_X\theta_{\rm BO}(X)|^2 + \lambda_0(X) = E\,. \tag{3.17}$$

3.3. **Equations for the density.** We note that

$$\phi = \hat G^{-1}\hat\psi = \left(\frac{\rho}{\langle\hat\psi,\hat\psi\rangle/\int\langle\hat\psi,\hat\psi\rangle\hat G^{-2}\,dX}\right)^{1/2}\hat\psi\,,$$

shows that $G$ and $\psi$ determine the density

$$\rho_{\rm S} = \rho = \frac{\langle\hat\psi,\hat\psi\rangle|\hat G|^{-2}}{\int\langle\hat\psi,\hat\psi\rangle\,|\hat G|^{-2}dX}\,, \tag{3.18}$$

defined in (3.14). Using the Born-Oppenheimer approximation in Lemma 8.2 we have $\langle\hat\psi,\hat\psi\rangle = 1+\mathcal{O}(M^{-1})$ in the case of a spectral gap. Therefore the weight function $|\hat G|^{-2}$ approximates the density and we know from Theorem 3.1 that $|\hat G|^{-2}$ is determined by the phase function $\theta$.

The Born-Oppenheimer dynamics generates an approximate solution $\Psi_{\rm BO}\hat G_{\rm BO}^{-1}e^{iM^{1/2}\theta_{\rm BO}}$ which yields the density

$$\rho_{\rm BO} = |\hat G_{\rm BO}|^{-2}, \tag{3.19}$$

where

$$\frac{d}{dt}\log|\hat G_{\rm BO}|^{-2} = -\Delta_X\theta_{\rm BO}(X)\,.$$

This representation can also be obtained from the conservation of mass

$$0 = \operatorname{div}(\rho_{\rm BO}\nabla_X\theta_{\rm BO}) \tag{3.20}$$

implying

$$\frac{d}{dt}\rho_{\rm BO}(X_t) = \nabla_X\rho_{\rm BO}(X_t)\cdot\dot X_t = -\rho_{\rm BO}(X_t)\,\operatorname{div}\nabla_X\theta_{\rm BO}\,, \tag{3.21}$$

with the solution

$$\rho_{\rm BO}(X_t) = \frac{C}{|\hat G_{\rm BO}(X_t)|^2}\,, \tag{3.22}$$

where $C$ is a positive constant for each characteristic. Note that the derivation of this classical density does not need a corresponding WKB equation but uses only the conservation of mass that holds for classical paths satisfying a Hamiltonian system. The classical density corresponds precisely to the Eulerian-Lagrangian change of coordinates $|G_t|^2/|G_0|^2 = \det(\partial X_t/\partial X_0)$ in (3.13).

3.4. **Construction of the solution operator.** The WKB Ansatz (3.1) is meaningful when $\phi$ does not include the full small scale. In Lemma 8.2 we present conditions for $\psi$ to be smooth.

To construct the solution operator it is convenient to include a non interacting particle in the system, i.e., a particle without charge, and assume that this particle moves with a constant, high speed $dX_1^1/dt = P_1^1 \gg 1$ (or equivalently with the unit speed and a large mass). Such a non interacting particle does not affect the other particles. The additional new coordinate $X_1^1$ is helpful in order to simply relate the time-coordinate $t$ and $X_1^1$. We add the corresponding kinetic energy $(P_1^1)^2/2$ to $E$ in order not to change the original problem (2.2) and write the equation (3.12) in the fast time scale $\tau = M^{1/2}t$

$$i\frac{d}{d\tau}\psi = (\mathcal{V} - V_0)\psi - \frac{1}{2M}G\sum_j \Delta_{X^j}(G^{-1}\psi)\,.$$

Furthermore, we change to the coordinates

$$(\tau, X_*) := (\tau, X_2^1, X_3^1, X^2, \ldots, X^N) \in [0,\infty)\times I\,,\ \text{ instead of } (X^1, X^2, \ldots, X^N) \in \mathbb{R}^{3N},$$

where $X^j = (X_1^j, X_2^j, X_3^j) \in \mathbb{R}^3$. Hence we obtain

$$i\dot{\psi} + \frac{1}{2(P_1^1)^2}\ddot{\psi} = (\mathcal{V} - V_0)\psi - \frac{1}{2M}G\sum_j \Delta_{X_*^j}(G^{-1}\psi) =: \tilde{\mathcal{V}}\psi\,, \tag{3.23}$$

using the notation $\dot{w} = dw/d\tau$ in this section. In Section 8.1 we show that the left hand side can be reduced to $i\dot{\psi}$ as $P_1^1 \to \infty$, by choosing special initial data. Note also that $G$ is independent of the first component in $X^1$. We see that the operator

$$\bar{\mathcal{V}} := G^{-1}\tilde{\mathcal{V}}G = \underbrace{G^{-1}(\mathcal{V} - V_0)G}_{=\mathcal{V}-V_0} - \frac{1}{2M}\sum_j \Delta_{X_*^j}$$

is symmetric on $L^2(\mathbb{R}^{3n+3N-1})$. Assume now the data $(X_0, P_0, Z_0)$ for $X_0 \in \mathbb{R}^{3N-1}$ is $(L\mathbb{Z})^{3N-1}$-periodic, then also $(X_\tau, P_\tau, Z_\tau)$ is $(L\mathbb{Z})^{3N-1}$-periodic, for $Z_t = \theta(X_t)$ and $P_t = \nabla_X\theta(X_t)$. To simplify the notation for such periodic functions, define the periodic circle

$$\mathbb{T} := \mathbb{R}/(L\mathbb{Z})\,.$$

We seek a solution $\Phi$ of (2.2) which is $(L\mathbb{Z})^{3(n+N)-1}$-periodic in the $(x, X_*)$-variable. The Schrödinger operator $\bar{\mathcal{V}}(\cdot, X_\tau)$ has, for each $\tau$, real eigenvalues $\{\lambda_m(\tau)\}$ with a complete set of eigenvectors $\{\zeta^m(x, X_*, \tau)\}$ orthogonal in the space functions, a subset of $L^2(\mathbb{T}^{3n+3N-1})$, see [1]. The proof uses that the operator $\bar{\mathcal{V}}_\tau + \gamma I$ generates a compact solution operator in the Hilbert space functions in $L^2(\mathbb{T}^{3n+3N-1})$, for the constant $\gamma \in (0,\infty)$ chosen sufficiently large. The discrete spectrum and the compactness comes from Fredholm theory for compact operators and the fact that the bilinear form $\int_{\mathbb{T}^{3(n+N)-1}} v\bar{\mathcal{V}}_\tau w + \gamma vw\,dx\,dX_*$ is continuous and coercive on $H^1(\mathbb{T}^{3(n+N)-1})$, see [12]. We see that $\tilde{\mathcal{V}}$ has the same eigenvalues $\{\lambda_m(\tau)\}$ and the eigenvectors $\{G_\tau\zeta^m(\tau)\}$, orthogonal in the weighted $L^2$-scalar product

$$\int_{\mathbb{T}^{3N-1}} \langle v, w\rangle\ \hat{G}^{-2}\,dX_*\,.$$

The construction and analysis of the solution operator continues in Section 8.1 based on the spectrum.

**Remark 3.2** (Boundary conditions)**.** The Schrödinger problem (2.2) makes sense not only in the periodic setting but also with alternative boundary conditions, e.g. from interaction with an external environment in scattering problems.

## 4. Fourier integral WKB states including caustics

4.1. **A preparatory example with the simplest caustic.** As an example of a caustic, we study first the simplest example of a fold caustic based on the Airy function $\mathrm{A} : \mathbb{R} \to \mathbb{R}$ which solves

$$-\partial_{xx}\mathrm{A}(x) + x\mathrm{A}(x) = 0\,. \tag{4.1}$$

The scaled Airy function

$$u(x) = C\,\mathrm{A}(M^{1/3}x)$$

solves the Schrödinger equation

$$-\frac{1}{M}\partial_{xx}u(x) + xu(x) = 0\,, \tag{4.2}$$

for any constant $C$. In our context an important property of the Airy function is the fact that it is the inverse Fourier transform of the function

$$\hat{\mathrm{A}}(p) = \sqrt{\frac{2}{\pi}}e^{ip^3/3}\,,$$

i.e.,

$$\mathrm{A}(x) = \frac{1}{\pi}\int_{\mathbb{R}} e^{i(xp+p^3/3)}\,dp\,. \tag{4.3}$$

In the next section, we will consider a general Schrödinger equation and determine a WKB Fourier integral corresponding to (4.3) for the Airy function; as an introduction to the general case we show how the derive (4.3): by taking the Fourier transform of the ordinary differential equation (4.1)

$$0 = \int_{\mathbb{R}} (-\partial_{xx} + x)\,\mathrm{A}(x)e^{-ixp}\,dx = (p^2 + i\partial_p)\hat{\mathrm{A}}(p)\,, \tag{4.4}$$

we obtain an ordinary differential equation for the Fourier transform $\hat{\mathrm{A}}(p)$ with the solution $\hat{\mathrm{A}}(p) = Ce^{ip^3}$, for any constant $C$. Then, by differentiation, it is clear that the scaled Airy function $u$ solves (4.2). Furthermore, the stationary phase method, cf. Section 10, shows that to the leading order $u$ is approximated by

$$u(x) \simeq C\left(-xM^{1/3}\right)^{-1/4}\cos\left(M^{1/2}(-x)^{3/2} - \pi/4\right), \quad \text{for } x < 0\,,$$

and $u(x) \simeq 0$ to any order (i.e., $\mathcal{O}(M^{-K})$ for any positive $K$) when $x > 0$. The behaviour of the Airy function is illustrated in Figure 3.

4.1.1. *Molecular dynamics for the Airy function.* The eikonal equation corresponding to (4.2) is

$$p^2 + x = 0$$

with solutions for $x \le 0$, which leads to the phase

$$p = \theta'(x) = \pm(-x)^{1/2}\,,\ \text{and}\ \ \theta(x) = \mp\frac{2}{3}(-x)^{3/2}\,. \tag{4.5}$$

We compute the Legendre transform

$$\theta^*(p) = xp - \theta(x)$$

where by (4.5) and $-x = p^2$ we obtain

$$\theta^*(p) = -p^2p + \frac{2}{3}p^3 = -\frac{p^3}{3}\,.$$

We note that this solution is also obtained from the eikonal equation

$$p^2 + \partial_p\theta^*(p) = 0\,,$$

which is solved by

$$\theta^*(p) = -p^3/3\,.$$

Thus we recover the relation for the Legendre transform $-xp + \theta^*(p) = -\theta(x)$.

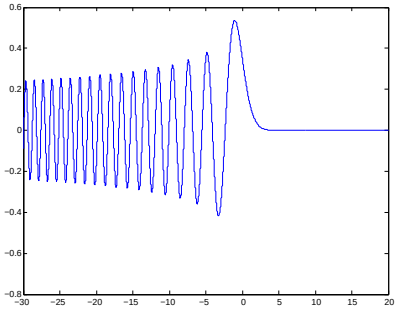


FIGURE 3. The Airy function.

4.1.2. *Observables for the Airy function.* The primary object of our analysis is an observable (a functional depending on $u$) rather than the solution $u(x)$ itself. Thus we first compute the observable evaluated on the solution obtained from the Airy function. In the following calculation we denote by $C$ a generic constant not necessarily the same at each occurrence,

$$
\begin{aligned}
\int_{\mathbb{R}} g(x)|u(x)|^2 dx &= C \int_{\mathbb{R}} g(x) \int_{\mathbb{R}}\int_{\mathbb{R}} e^{-iM^{1/2}(xp+p^3/3)} e^{iM^{1/2}(xq+q^3/3)}\, dq\, dp\, dx \\
&= C \int_{\mathbb{R}}\int_{\mathbb{R}} \hat{g}\left(M^{1/2}(p-q)\right) e^{iM^{1/2}(q^3/3-p^3/3)}\, dq\, dp \\
&= C \int_{\mathbb{R}}\int_{\mathbb{R}} \hat{g}\left(M^{1/2}(p-q)\right) e^{iM^{1/2}\left((q-p)^3/12+(q-p)(p+q)^2/4\right)}\, dq\, dp \\
&= C \int_{\mathbb{R}}\int_{\mathbb{R}} \hat{g}(\underbrace{-M^{1/2}\bar{q}}_{=t})\, e^{iM^{1/2}\left(\bar{q}^3/12+\bar{q}\,\bar{p}^2/4\right)} \overbrace{d\bar{q}d\bar{p}}^{\bar{q}=q-p,\,\bar{p}=p+q} \\
&= C \int_{\mathbb{R}}\int_{\mathbb{R}} \hat{g}(t) e^{-i\left(t^3/(12M)+t\,\bar{p}^2/4\right)}\, dt\, d\bar{p} \\
&= C \int_{\mathbb{R}} g * \mathrm{A}_M(\underbrace{-\bar{p}^2}_{=\partial_p\theta^*(\bar{p})=x})\, d\bar{p} \\
&= C \int_{-\infty}^{0} g * \mathrm{A}_M(x)|\partial_x p(x)|\, dx\,,
\end{aligned}
\tag{4.6}
$$

where

$$
\mathrm{A}_M(x) := \left(\frac{M}{4}\right)^{1/3} \mathrm{A}\left(\left(\frac{M}{4}\right)^{1/3} x\right) \text{ is the Fourier transform of } e^{-it^3/(12M)}. \tag{4.7}
$$

**Lemma 4.1.** *The scaled Airy function* $\mathrm{A}_M$ *is an approximate identity in the following sense*

$$\|g * \mathrm{A}_M - g\|_{L^2(\mathbb{R})} \leq \frac{1}{12M} \|\partial_x^3 g\|_{L^2(\mathbb{R})} \,. \tag{4.8}$$

*Proof.* Plancherel's Theorem implies

$$\begin{aligned} M\|g * \mathrm{A}_M - g\|_{L^2} &= M\|\hat{g}\hat{\mathrm{A}}_M - \hat{g}\|_{L^2} = \|\hat{g}(e^{ip^3/(12M)} - 1)M\|_{L^2} \\ &\leq \frac{1}{12}\||p|^3\hat{g}\|_{L^2} = \frac{1}{12}\|\partial_x^3 g\|_{L^2} \,. \end{aligned}$$

The inequality follows from $|e^{iy} - 1| \leq |y|$ which holds for all $y \in \mathbb{R}$. □

The classical molecular dynamics approximation corresponding to the Schrödinger equation (4.2) is the Hamiltonian system

$$\dot{X} = p \,, \quad \dot{p} = -\frac{1}{2}$$

with a solution $X_t = -t^2/4$ and the corresponding approximation of the observable

$$\frac{1}{T}\int_0^T g(X_t)\, dt = \frac{1}{T}\int_0^T g(X_t)\, \frac{dX_t}{\dot{X}_t} = \frac{1}{T}\int_{-T^2/4}^0 g(x)\, \frac{dx}{|p(x)|} \,.$$

In this specific case the phase satisfies $|p(x)| = |x|^{1/2}$ and $|\partial_x p| = |x|^{-1/2}/2$, and hence the non-normalized density $|p|^{-1}$ is in this case equal to $2|\partial_x p|$. Equation (4.6) and Lemma 4.1 imply

$$\Big|\int_{\mathbb{R}} g|u|^2\, dx - \int_{\mathbb{R}} g\partial_x p(x)\, dx\Big| = \mathcal{O}(M^{-1})$$

and consequently for two different observables $g_1$ and $g_2$ we have that Schrödinger observables are approximated by the classical observables with the error $\mathcal{O}(M^{-1})$

$$\frac{\int_{\mathbb{R}} g_1|u|^2\, dx}{\int_{\mathbb{R}} g_2|u|^2\, dx} - \frac{\int_{\mathbb{R}} g_1|\partial_{xx}\theta|\, dx}{\int_{\mathbb{R}} g_2|\partial_{xx}\theta|\, dx} = \mathcal{O}(M^{-1}) \,, \tag{4.9}$$

using $\partial_x p(x) = \partial_{xx}\theta(x)$. The reason we compare two different observables with a compact support is that $\int_{\mathbb{R}} u^2(x)\, dx = \infty$ in the case of the Airy function.

We note that in (4.6) we used

$$\frac{1}{3}(q^3 - p^3) = \theta^*(p) - \theta^*(q) = (p-q)\partial_p\theta^*\left(\frac{1}{2}(p+q)\right) + \frac{1}{3}\partial^3\theta^*\left(\frac{1}{2}(p+q)\right)\left(\frac{1}{2}(p-q)\right)^3$$

which in the next section is generalized to other caustics. For the Airy function there holds

$$\frac{1}{3}\partial^3\theta^*\left(\frac{1}{2}(p+q)\right) = -\frac{2}{3} \,.$$

4.2. **A general Fourier integral ansatz.** In order to treat a more general case with a caustic of the dimension $d$ we use the Fourier integral ansatz

$$\Phi(X, x) = \int_{\mathbb{R}^d} \phi(X, x) e^{-iM^{1/2}\Theta(\check{X}, \hat{X}, \check{P})}\, d\check{P} \tag{4.10}$$

and we write

$$\begin{aligned} X = (\hat{X}, \check{X}) \,, \quad P = (\hat{P}, \check{P}) \\ \check{X}\cdot\check{P} = \sum_{j=1}^{d} \check{X}^j\check{P}^j \,, \quad \hat{X}\cdot\hat{P} = \sum_{j=d+1}^{N} \hat{X}^j\hat{P}^j \\ \Theta(\check{X}, \hat{X}, \check{P}) = \check{X}\cdot\check{P} - \theta^*(\hat{X}, \check{P}) \,, \end{aligned}$$

based on the Legendre transform

$$\theta^*(\hat{X}, \check{P}) = \min_{\check{X}} \left(\check{X}\cdot\check{P} - \theta(\hat{X}, \check{X})\right) \,.$$

If the function $\theta^*(\hat{X},\check{P})$ is not defined for all $\check{P}\in\mathbb{R}^d$, but only for $\check{P}\in\mathcal{U}\subset\mathbb{R}^d$ we replace the integral over $\mathbb{R}^d$ by integration over $\mathcal{U}$ using a smooth cut-off function $\chi(\check{P})$. The cut-off function is zero outside $\mathcal{U}$ and equal to one in a large part of the interior of $\mathcal{U}$, see Section 4.2.3. The ansatz (4.10) is inspired by Maslov's work [25], although it is not the same since our amplitude function $\phi$ depends on $(\hat{X},\check{X},x)$ but not on $\check{P}$. We emphasize that our modification consisting in having an amplitude function that is not dependent on $\check{P}$ is essential in the construction of the solution and for determining the accuracy of observables based on this solution.

4.2.1. *Making the ansatz for a Schrödinger solution.* In this section we construct a solution to the Schrödinger equation from the ansatz (4.10). The constructed solution will be an *actual* solution and not only an asymptotic solution as in [25]. We consider first the case when the integration is over $\mathbb{R}^d$ and then conclude in the end that the cut-off function $\chi(\check{P})$ can be included in all integrals without changing the property of the Fourier integral ansatz being a solution in the $\check{X}$-domain where $\check{X}=\nabla_{\check{P}}\theta^*(\hat{X},\check{P})$ for some $\check{P}$ satisfying $\chi(\check{P})=1$.

The requirement to be a solution means that there should hold

(4.11)
$$\begin{aligned}0&=(\mathcal{H}-E)\Phi\\&=\int_{\mathbb{R}^d}\left(\frac{1}{2}|\nabla_{\hat{X}}\theta^*(\hat{X},\check{P})|^2+\frac{1}{2}|\check{P}|^2+V_0(X)-E\right)\phi(X,x)e^{-iM^{1/2}\Theta(\check{X},\hat{X},\check{P})}\,d\check{P}\\&\quad-\int_{\mathbb{R}^d}\left(iM^{-1/2}(\nabla_{\hat{X}}\phi\cdot\nabla_{\hat{X}}\theta^*-\nabla_{\check{X}}\phi\cdot\check{P}+\frac{1}{2}\phi\Delta_{\hat{X}}\theta^*)-(\mathcal{V}-V_0)\phi+\frac{1}{2M}\Delta_X\phi\right)e^{-iM^{1/2}\Theta(\check{X},\hat{X},\check{P})}\,d\check{P}\,.\end{aligned}$$

Comparing this expression to the previously discussed case of a single WKB-mode we see that the zero order term is now $\Delta_{\hat{X}}\theta^*$ instead of $\Delta_X\theta$ and that we have $-\nabla_{\check{X}}\phi\cdot\check{P}$ instead of $\nabla_{\check{X}}\phi\cdot\nabla_{\check{X}}\theta$. However, the main difference is that the first integral is not zero (only the leading order term of its stationary phase expansion is zero, cf. (10.1)). Therefore, the first integral contributes to the second integral. The goal is now to determine a function $F(\hat{X},\check{X},\check{P})$ satisfying

(4.12)
$$\begin{aligned}&\int_{\mathbb{R}^d}\left(\frac{1}{2}|\nabla_{\hat{X}}\theta^*|^2+\frac{1}{2}|\check{P}|^2+V_0(X)-E\right)e^{-iM^{1/2}\Theta(\check{X},\hat{X},\check{P})}\,d\check{P}\\&\qquad=iM^{-1/2}\int_{\mathbb{R}^d}F(\hat{X},\check{X},\check{P})\,e^{-iM^{1/2}\Theta(\check{X},\hat{X},\check{P})}\,d\check{P}\,,\end{aligned}$$

and verify that it is bounded.

**Lemma 4.2.** *There holds $F=F_0+F_1$ where*

$$F_0=\frac{1}{2}\sum_{i,j}\partial_{\check{X}^i\check{X}^j}V_0\left(\nabla_{\check{P}}\theta^*(\check{P})\right)\partial_{\check{P}^j\check{P}^i}\theta^*(\check{P})\,,$$

$$F_1=iM^{-1/2}\int_0^1\int_0^1\int_{\mathbb{R}^d}\sum_{i,j,k}t(1-t)\partial_{\check{P}^k}\left[\partial_{\check{X}^i\check{X}^j\check{X}^k}V_0\left(\nabla_{\check{P}}\theta^*(\check{P})+s\,t\,\delta\theta^*(\check{P})\right)\partial_{\check{P}_j\check{P}_i}\nabla_{\check{P}}\theta^*(\check{P})\right]\,dt\,ds\,.$$

*Proof.* The function $\theta^*(\hat{X},\check{P})$ is defined as a solution to the Hamilton-Jacobi (eikonal) equation

$$\frac{1}{2}|\nabla_{\hat{X}}\theta^*(\hat{X},\check{P})|^2+\frac{1}{2}|\check{P}|^2+V_0\left(\hat{X},\nabla_{\check{P}}\theta^*(\hat{X},\check{P})\right)-E=0 \tag{4.13}$$

for all $(\hat{X},\check{P})$. Consequently, the integral on the left hand side of (4.12) is

$$\int_{\mathbb{R}^d}\left(V_0(\hat{X},\check{X})-V_0(\hat{X},\nabla_{\check{P}}\theta^*(\hat{X},\check{P})\right)e^{-iM^{1/2}\left(\check{X}\cdot\check{P}-\theta^*(\hat{X},\check{P})\right)}\,d\check{P}\,.$$

Let $\check{P}_0(\check{X})$ be any solution to the stationary phase equation $\check{X}=\nabla_{\check{P}}\theta^*(\hat{X},\check{P}_0)$ and introduce the notation

$$\Theta'(\check{X},\hat{X},\check{P}):=\nabla_{\check{P}}\theta^*(\hat{X},\check{P}_0)\cdot\check{P}-\theta^*(\hat{X},\check{P})\,.$$

Then by writing a difference as $V(y_1)-V(y_2)=\int_0^1 \partial_y V(y_2+t(y_1-y_2))dt\cdot(y_1-y_2)$, identifying a derivative $\partial_{\check P_i}$ and integrating by parts the integral can be written

$$
\begin{aligned}
&\int_{\mathbb{R}^d}\Big(V_0(\hat X,\nabla_{\check P}\theta^*(\hat X,\check P_0))-V_0(\hat X,\nabla_{\check P}\theta^*(\hat X,\check P)\Big)\,e^{-iM^{1/2}\Theta'(\check X,\hat X,\check P)}\,d\check P\\
&=\int_0^1\int_{\mathbb{R}^d}\sum_i \partial_{\check X^i}V_0\left(\nabla_{\check P}\theta^*(\check P)+t\,\left[\nabla_{\check P}\theta^*(\check P_0)-\nabla_{\check P}\theta^*(\check P)\right]\right)\times\\
&\qquad\times\left(\partial_{\check P^i}\theta^*(\check P_0)-\partial_{\check P^i}\theta^*(\check P)\right)e^{-iM^{1/2}\Theta'(\check X,\hat X,\check P)}\,d\check P\,dt\\
&=-iM^{-1/2}\int_0^1\int_{\mathbb{R}^d}\sum_i \partial_{\check X^i}V_0\left(\nabla_{\check P}\theta^*(\check P)+t\,\left[\nabla_{\check P}\theta^*(\check P_0)-\nabla_{\check P}\theta^*(\check P)\right]\right)\partial_{\check P_i}e^{-iM^{1/2}\Theta'(\check X,\hat X,\check P)}\,d\check P\,dt\\
&=iM^{-1/2}\int_0^1\int_{\mathbb{R}^d}\sum_i \partial_{\check P_i}\partial_{\check X^i}V_0\left(\nabla_{\check P}\theta^*(\check P)+t\,\left[\nabla_{\check P}\theta^*(\check P_0)-\nabla_{\check P}\theta^*(\check P)\right]\right)e^{-iM^{1/2}\Theta'(\check X,\hat X,\check P)}\,d\check P\,dt\,.
\end{aligned}
$$

Therefore the leading order term in $F=:F_0+F_1$ is

$$
\begin{aligned}
F_0&:=\int_0^1\sum_{i,j}(1-t)\partial_{\check X^i\check X^j}V_0\left(\nabla_{\check P}\theta^*(\check P)\right)\partial_{\check P^j\check P^i}\theta^*(\check P)\,dt\\
&=\frac12\sum_{i,j}\partial_{\check X^i\check X^j}V_0\left(\nabla_{\check P}\theta^*(\check P)\right)\partial_{\check P^j\check P^i}\theta^*(\check P)\,.
\end{aligned}
$$

Denoting $\delta\theta^*(\check P)=\nabla_{\check P}\theta^*(\check P_0)-\nabla_{\check P}\theta^*(\check P)$ the remainder becomes

$$
\begin{aligned}
&-iM^{-1/2}\int_0^1\int_{\mathbb{R}^d}\sum_{i,j}\left[\partial_{\check X^i\check X^j}V_0\left(\nabla_{\check P}\theta^*(\check P)\right)-\partial_{\check X^i\check X^j}V_0\left(\nabla_{\check P}\theta^*(\check P)+t\,\delta\theta^*(\check P)\right)\right]\\
&\qquad\times(1-t)\partial_{\check P^j\check P^i}\theta^*(\check P)\,e^{-iM^{1/2}\Theta'(\check X,\hat X,\check P)}\,d\check P\,dt\\
&=iM^{-1/2}\int_0^1\int_0^1\int_{\mathbb{R}^d}\sum_{i,j,k}t(1-t)\partial_{\check X^i\check X^j\check X^k}V_0\left(\nabla_{\check P}\theta^*(\check P)+s\,t\,\delta\theta^*(\check P)\right)\partial_{\check P^j\check P^i}\theta^*(\check P)\\
&\qquad\times\left(\partial_{\check P^k}\theta^*(\check P_0)-\partial_{\check P^k}\theta^*(\check P)\right)e^{-iM^{1/2}\Theta'(\check X,\hat X,\check P)}\,d\check P\,dt\,ds\\
&=-\frac{1}{M}\int_0^1\int_0^1\int_{\mathbb{R}^d}\sum_{i,j,k}t(1-t)\partial_{\check P^k}\left[\partial_{\check X^i\check X^j\check X^k}V_0\left(\nabla_{\check P}\theta^*(\check P)+s\,t\,\delta\theta^*(\check P)\right)\partial_{\check P^j\check P^i}\theta^*(\check P)\right]\\
&\qquad\times e^{-iM^{1/2}\Theta'(\check X,\hat X,\check P)}\,d\check P\,dt\,ds\,,
\end{aligned}
$$

hence the function $F_1$ is purely imaginary and small

$$
F_1=iM^{-1/2}\int_0^1\int_0^1\int_{\mathbb{R}^d}\sum_{i,j,k}t(1-t)\partial_{\check P^k}\left[\partial_{\check X^i\check X^j\check X^k}V_0\left(\nabla_{\check P}\theta^*(\check P)+s\,t\,\delta\theta^*(\check P)\right)\partial_{\check P_j\check P_i}\nabla_{\check P}\theta^*(\check P)\right]\,dt\,ds\,,
$$

and

$$
2\mathrm{Re}\,F=\sum_{i,j}\partial_{\check X^i\check X^j}V_0\left(\nabla_{\check P}\theta^*(\check P)\right)\partial_{\check P^j\check P^i}\theta^*(\check P)\,. \tag{4.14}
$$

□

The eikonal equation (4.13) and the requirement that $(\mathcal{H}-E)\Phi=0$ in (4.11) then imply that

$$
\begin{aligned}
0=\int_{\mathbb{R}^d}\Bigg[&iM^{-1/2}\left(\nabla_{\hat X}\phi\cdot\nabla_{\hat X}\theta^*-\nabla_{\check X}\phi\cdot\check P+\frac12\phi\left(\Delta_{\hat X}\theta^*-2F(X,\check P)\right)\right)\\
&-(\mathcal{V}-V_0)\phi+\frac{1}{2M}\Delta_X\phi\Bigg]\,e^{-iM^{1/2}\Theta(\check X,\hat X,\check P)}\,d\check P\,.
\end{aligned} \tag{4.15}
$$

The Hamilton-Jacobi eikonal equation (4.13), in the primal variable $(\hat{X}, \check{P})$ with the corresponding dual variable $\hat{P}, \check{X})$, can be solved by the characteristics

$$\begin{aligned} \dot{\hat{X}} &= \hat{P} \\ \dot{\hat{P}} &= -\nabla_{\hat{X}} V_0(\hat{X}, \check{X}) \\ \dot{\check{X}} &= -\check{P} \\ \dot{\check{P}} &= \nabla_{\check{X}} V_0(\hat{X}, \check{X}) \,, \end{aligned} \tag{4.16}$$

using the definition

$$\begin{aligned} \nabla_{\hat{X}} \theta^*(\hat{X}, \check{P}) &= \hat{P} \\ \nabla_{\check{P}} \theta^*(\hat{X}, \check{P}) &= \check{X} \,. \end{aligned}$$

The characteristics give

$$\frac{d}{dt}\phi = \nabla_{\hat{X}}\phi \cdot \nabla_{\hat{X}}\theta^* - \nabla_{\check{X}}\phi \cdot \check{P} \,,$$

so that the Schrödinger transport equation becomes, as in (3.12),

$$iM^{-1/2}\left(\dot{\phi} + \phi\frac{\dot{G}}{G}\right) = (\mathcal{V} - V_0)\phi - \frac{1}{2M}\Delta_X\phi \tag{4.17}$$

and for $\psi = G\phi$

$$iM^{-1/2}\dot{\psi} = (\mathcal{V} - V_0)\psi - \frac{G}{2M}\Delta_X\frac{\psi}{G} \tag{4.18}$$

with the complex valued weight function $G$ defined by

$$\frac{d}{dt}\log G_t = \frac{1}{2}\Delta_{\hat{X}}\theta^*(\hat{X}_t, \check{P}_t) - F(\hat{X}_t, \check{P}_t) \,. \tag{4.19}$$

This transport equation is of the same form as the transport equation for a single WKB-mode, with a modification of the weight function $G$.

Differentiation of the second equation in the Hamiltonian system (4.16) implies that the first variation $\partial\check{P}_t/\partial\check{X}_0$ satisfies

$$\frac{d}{dt}\left(\frac{\partial\check{P}_t^i}{\partial\check{X}_0}\right) = \sum_{j,k}\partial_{\check{X}^i\check{X}^j}V_0(\hat{X}, \check{X}_t)\partial_{\check{P}^j\check{P}^k}\theta^*(\check{P})\frac{\partial\check{P}_t^k}{\partial\check{X}_0} \,,$$

which by the Liouville formula (3.13) and the equality

$$2\mathrm{Re}\, F = \sum_{i,j}\partial_{\check{X}^i\check{X}^j}V_0\partial_{\check{P}^j\check{P}^i}\theta^* = \mathrm{Tr}\,(\sum_j \partial_{\check{X}^i\check{X}^j}V_0\partial_{\check{P}^j\check{P}^k}\theta^*)$$

in (4.14) yields the relation,

$$e^{-2\int_0^t \mathrm{Re}\, F\, dt'} = C\left|\det\frac{\partial\check{P}_t}{\partial\check{X}_0}\right| \,, \tag{4.20}$$

for the constant $C := |\det\frac{\partial\check{X}_0}{\partial\check{P}_0}|$. We use relation (4.20) to study the density in the next section.

**Remark 4.3.** The conclusion in this section holds also when all integrals over $d\check{P}$ in $\mathbb{R}^d$ are replaced by integrals with the measure $\chi(\check{P})\,d\check{P}$. Then there holds $2\mathrm{Re}\, F = \sum_{ij}\partial_{\check{X}^i\check{X}^j}\mathcal{V}\partial_{\check{P}^i}(\chi\partial_{\check{P}^j}\theta^*)$. We use that the observable $g$ is zero when the cut-off function $\chi_j$ is not one, see Section 4.2.3. In Section 5 we show how to construct a global solution by connecting the Fourier integral solutions, valid in a neighborhood where $\det\partial(X)/\partial(P)$ vanishes (and $\chi(\check{P}) = 1$), to a sum of WKB-modes, valid in neighborhoods where $\det\partial(P)/\partial(X)$ vanishes (and $\chi(\check{P}) < 1$).

4.2.2. *The Schrödinger density for caustics.* In this section we study the density generated by the solution

$$\Phi(X,x)=\int_{\mathbb{R}^d}\phi(X,x)\,e^{-iM^{1/2}\left(\check{X}\cdot\check{P}-\theta^*(\hat{X},\check{P})\right)}\,d\check{P}\,.$$

The analysis of the density generalizes the calculations for the Airy function in Section 4.1.2. We have, using the notation $\hat{\tilde{g}}$ for the Fourier transform of $\tilde{g}$ with respect to the $\check{X}$ variable, and by introducing the notation $\check{R}=\frac{1}{2}(\check{P}+\check{Q})$ and $\check{S}=\check{P}-\check{Q}$

$$\begin{aligned}
\int g(X)|\Phi(x,X)|^2\,dxdX &= \int \underbrace{g(X)\langle\phi,\phi\rangle}_{=:\tilde{g}(X)}\,e^{iM^{1/2}(\check{X}\cdot\check{P}-\theta^*(\hat{X},\check{P}))}\,e^{-iM^{1/2}(\check{X}\cdot\check{Q}-\theta^*(\hat{X},\check{Q}))}\,d\check{P}\,d\check{Q}\,dX\\
&= \int \hat{\tilde{g}}(\hat{X},M^{1/2}\check{S})\,e^{iM^{1/2}(\theta^*(\hat{X},\check{Q})-\theta^*(\hat{X},\check{P}))}\,d\check{P}\,d\check{Q}\,d\hat{X}\\
&= \int \hat{\tilde{g}}(\hat{X},M^{1/2}\check{S})\,e^{iM^{1/2}\frac{1}{6}(\check{S}\cdot\nabla_{\check{P}})^3\theta^*(\hat{X},\check{R}+\gamma\check{S}/2)}\times\\
&\qquad\times e^{iM^{1/2}\check{S}\cdot\nabla_{\check{P}}\theta^*(\hat{X},\check{R})}\,d\check{S}\,d\check{R}\,d\hat{X}\\
&= \left(\frac{1}{2\pi}\right)^{d/2}M^{-1/2}\int \tilde{g}*\mathrm{A}_M\big(\hat{X},\underbrace{\nabla_{\check{P}}\theta^*(\hat{X},\check{R})}_{=\check{X}}\big)\,d\check{R}d\hat{X}\\
&= \left(\frac{1}{2\pi}\right)^{d/2}M^{-1/2}\int \tilde{g}*\mathrm{A}_M(\hat{X},\check{X})\,\left|\det\frac{\partial(\check{P})}{\partial(\check{X})}\right|\,dX\,.
\end{aligned}\tag{4.21}$$

In the convolution $\tilde{g}*\mathrm{A}_M$, the function $\mathrm{A}_M$, analogous to (4.7), is the Fourier transform of

$$e^{i\frac{1}{M}(\omega\cdot\nabla_{\check{P}})^3\theta^*(\hat{X},\check{P})}\Big|_{\check{P}=\check{R}+\gamma\omega}$$

with respect to $\omega\in\mathbb{R}^d$ and the integration in $\check{X}$ is with respect to the range of $\nabla_{\check{P}}\theta^*(\hat{X},\cdot)$. As a next step we evaluate the Fourier transform and its derivatives at zero and obtain

$$\begin{aligned}
&\int_{\mathbb{R}^d}\mathrm{A}_M(\check{X})\,d\check{X}=1\,,\quad \int_{\mathbb{R}^d}\check{X}^i\mathrm{A}_M(\check{X})\,d\check{X}=0\,,\\
&\int_{\mathbb{R}^d}\check{X}^i\check{X}^j\mathrm{A}_M(\check{X})\,d\check{X}=0\,,\quad M\int_{\mathbb{R}^d}\check{X}^i\check{X}^j\check{X}^k\mathrm{A}_M(\check{X})\,d\check{X}=\mathcal{O}(1).
\end{aligned}$$

Here we use that both differentiation with respect to $(\omega\cdot\nabla_{\check{P}})^3$ and $\theta^*(\hat{X},\check{R}+\gamma\omega)$ yield factors of $\omega$ which vanish. The vanishing moments of $\mathrm{A}_M$ imply that

$$\|\tilde{g}*\mathrm{A}_M-\tilde{g}\|_{L^2(d\check{X})}=\mathcal{O}(M^{-1})\tag{4.22}$$

as in (4.8), so that up to $\mathcal{O}(M^{-1})$ error the convolution with $\mathrm{A}_M$ can be neglected.

4.2.3. *Integration over a compact set in $\check{P}$.* In the case when the integration is over $\mathcal{U}\subset\mathbb{R}^d$ instead of $\mathbb{R}^d$, we use a smooth cut-off function $\chi(\check{P})$, which is zero outside $\mathcal{U}$ and restrict our analysis to the case when the smooth observable mapping $\check{P}\mapsto g(\hat{X},\nabla_{\check{P}}\theta^*(\hat{X},\check{P}))$ is compactly supported in the domain where $\chi$ is one. In this way $g(\hat{X},\nabla_{\check{P}}\theta^*(\hat{X},\check{P}))$ is zero when $\nabla_{\check{P}}\chi(\check{P})$ is non zero. The integrand is thus equal to

$$(g(X)\ \langle\phi,\phi\rangle)\chi(\check{P})\chi(\check{Q})$$

and we use the convergent Taylor expansion

$$\chi(\underbrace{\check{R}+M^{-1/2}\omega}_{\check{P}})\chi(\underbrace{\check{R}-M^{-1/2}\omega}_{\check{Q}})=\sum_{k=0}^{\infty}\frac{|\omega|^{2k}}{M^k}a_k(\check{R})\,.$$

Then the observable becomes

$$(2\pi)^{-d/2}M^{-1/2}\sum_{k=0}^{\infty}\int\left(a_k\,(M^{-1}\Delta_{\check{X}})^k\tilde{g}\right)*\mathrm{A}_M\left(\hat{X},\nabla_{\check{P}}\theta^*(\hat{X},\check{R})\right)\,d\check{R}\,d\hat{X}\,.$$

As in (4.22) we can remove the convolution with $\mathrm{A}_M$ by introducing an error $\mathcal{O}(M^{-1})$ and since for $k>0$ we have $a_k(\check{R})g(\hat{X},\nabla_{\check{P}}\theta^*(\hat{X},\check{R}))=0$ and $a_0=1$, we obtain the same observable as before

$$\begin{aligned}
&\sum_{k=0}^{\infty}\int \left(a_k\,(M^{-1}\Delta_{\check{X}})^k\tilde{g}\right) * \mathrm{A}_M\big(\hat{X},\nabla_{\check{P}}\theta^*(\hat{X},\check{R})\big)\,d\check{R}\,d\hat{X}\\
&=\sum_{k=0}^{\infty}\int \left(a_k\,(M^{-1}\Delta_{\check{X}})^k\tilde{g}\right)\Big(\hat{X},\nabla_{\check{P}}\theta^*(\hat{X},\check{R})\Big)\,d\check{R}\,d\hat{X}\;+\mathcal{O}(M^{-1})\\
&=\int \tilde{g}\Big(\hat{X},\nabla_{\check{P}}\theta^*(\hat{X},\check{R})\Big)\,d\check{R}\,d\hat{X}\;+\mathcal{O}(M^{-1})\\
&=\int \tilde{g}(\hat{X},\check{X})\left|\det\frac{\partial(\check{P})}{\partial(\check{X})}\right|\,dX\;+\mathcal{O}(M^{-1})\,.
\end{aligned}$$

4.2.4. *Comparing the Schrödinger and molecular dynamics densities.* We compare the Schrödinger density to the molecular dynamics density generated by the continuity equation

$$0=\operatorname{div}(\rho\nabla\theta)=\nabla\rho\cdot\nabla\theta+\rho\operatorname{div}(\nabla\theta)=\dot{\rho}+\rho\operatorname{div}(\nabla\theta)$$

which yields the density

$$e^{-\int\operatorname{div}(\nabla\theta)\,dt}\,.$$

We have $P=\nabla\theta$, so that $\frac{\partial(P)}{\partial(X)}=\partial_{XX}\theta$. The Liouville formula (3.13) implies the molecular dynamics density

$$\rho_{\mathrm{BO}}=e^{-\int_0^t\operatorname{div}(\nabla\theta)\,dt'}=\det\frac{\partial X_{0,\mathrm{BO}}}{\partial X_{t,\mathrm{BO}}}\,. \tag{4.23}$$

The observable for the Schrödinger equation has, by (4.21), the density

$$(g\langle\phi,\phi\rangle)*\mathrm{A}_M\left|\det\frac{\partial(\check{P})}{\partial(\check{X})}\right|\,.$$

We want to compare it with the molecular dynamics density $\rho_{\mathrm{BO}}$. The convolution with $\mathrm{A}_M$ gives an error term of the order $\mathcal{O}(M^{-1})$, as in (4.8). The Schrödinger transport equation (4.17) and the definition of the weight $G$ in (4.19), show that the amplitude function satisfies, by (4.18) and (4.19) and the Born-Oppenheimer approximation Lemma 8.2,

$$\langle\phi,\phi\rangle=|G|^2\langle\psi,\psi\rangle=e^{\int 2\mathrm{Re}\;F-\Delta_{\hat{X}}\theta^*\,dt}+\mathcal{O}(M^{-1}),$$

so that by (4.20)

$$\begin{aligned}
(g\langle\phi,\phi\rangle)*\mathrm{A}_M|\det\frac{\partial(\check{P})}{\partial(\check{X})}| &= (g\langle\phi,\phi\rangle)|\det\frac{\partial(\check{P})}{\partial(\check{X})}|+\mathcal{O}(M^{-1})\\
&= g\,e^{\int 2\mathrm{Re}\,F-\Delta_{\hat{X}}\theta^*\,dt}\,|\det\frac{\partial(\check{P})}{\partial(\check{X})}|+\mathcal{O}(M^{-1})\\
&= g\,|\det\frac{\partial(\check{X}_0)}{\partial(\check{P})}|\,|\det\frac{\partial(\hat{X}_0)}{\partial(\hat{X})}|\,|\det\frac{\partial(\check{P})}{\partial(\check{X})}|+\mathcal{O}(M^{-1})\,,\\
&= g\,|\det\frac{\partial(\check{X}_0)}{\partial(\check{X})}|\,|\det\frac{\partial(\hat{X}_0)}{\partial(\hat{X})}|+\mathcal{O}(M^{-1})\,,\\
&= g\,|\det\frac{\partial(X_0)}{\partial(X)}|+\mathcal{O}(M^{-1})\,.
\end{aligned} \tag{4.24}$$

When we restrict the domain to $\mathcal{U}$ with the cut-off function $\chi$ as in Remark 4.3 we use the fact that $g(\hat{X},\nabla_{\check{P}}\theta^*(\hat{X},\check{P}))$ is zero when $\nabla_{\check{P}}\chi(\check{P})$ is non zero and obtain the same. The representations (4.24) and (4.23) show that the density generated in the caustic case with a Fourier integral also takes the same form, to the leading order, as the molecular dynamics density and the remaining discrepancy is only due to $\theta^*=\theta^*_{\mathrm{S}}$ and $\theta^*=\theta^*_{\mathrm{BO}}$ being different. This difference is, as in the single mode WKB expansion, of size $\mathcal{O}(M^{-1})$ which is estimated by the difference in Hamiltonians of the Schrödinger and molecular dynamics eikonal

equations. The estimate of the difference of the phase functions uses the Hamilton-Jacobi equation (4.13) for $\theta_{\rm S}^*(\hat{X},\check{P})$ and a similar Hamilton-Jacobi equation for $\theta_{\rm BO}^*(\hat{X},\check{P})$ with $V_0 = \lambda_{\rm BO} + \mathcal{O}(M^{-1})$ replaced by $\lambda_{\rm BO}$. The difference in the weight functions $\log(|G(\hat{X},\check{P})|^{-2})$ is estimated by the Hamilton-Jacobi equation

$$\left(\nabla_{\hat{X}}\theta_{\rm S}^*(\hat{X},\check{P})\cdot\nabla_{\hat{X}} - \nabla_{\check{X}}V_0(\hat{X},\check{X})\cdot\nabla_{\check{P}}\right)\log|G_{\rm S}(\hat{X},\check{P})|^{-2} - \Delta_{\hat{X}}\theta_{\rm S}^*(X,\check{P}) + \mathrm{Re}\,F(X,\check{P}) = 0\,,$$

where $\mathrm{Re}\,F$ is given in (4.14), and by the similar Hamilton-Jacobi equation with $V_0 = \lambda_{\rm BO} + \mathcal{O}(M^{-1})$ replaced by $\lambda_{\rm BO}$ and $\theta_{\rm S}^*$ by $\theta_{\rm BO}^*$.

## 5. A global construction coupling caustics with single WKB-modes

We use the paths generated by the Hamiltonian to construct a global asymptotic solution. More precisely, we will construct a $3N$-dimensional Lagrangian manifold from a Hamiltonian; for this we need some initial data, i.e. a $3N-1$ dimensional Lagrangian manifold, which we construct from a Poincare map in four steps. The basic properties of a Lagrangian manifold we use are reviewed in Section 5.1 and the basic assumption we make is that the Lagrangian manifold is smooth, which excludes ergodic dynamics where the Lagrangian manifold is dense in set of dimension $6N-1$.

**Step 1. Define a hitting plane.** Consider a codimension one hitting plane in the $X-$coordinate space, e.g. $X_{1_1} = 0$.

**Step 2. Consider a WKB Schrödinger solution and its initial Lagrangian manifold data.** We seek initial data on the hitting plane $X_{1_1} = 0$ in the form of a (smooth) $3N-1$ dimensional Lagrangian manifold $L$, satisfying the two constraints $H(X,P) = E$ & $X_{1_1} = 0$ for $(X,P) \in L$. For instance any smooth function $G : \mathbb{R}^{3N} \to \mathbb{R}$ generates a local subset of a Lagrangian manifold

$$L \supseteq \{(X,P(X)) \mid \text{ for all } X \text{ such that } H(X,P(X)) = E, \,\&\, X_{1_1} = 0,\ P(X) = \nabla_X G\}$$

and one can permute the role of $X$ and $P$ to obtain other parts of the set. Section 3.1.1 shows, in the case of piecewise constant potentials, that the solution to the Schrödinger equation is of the form $\Phi(X) = \sum_j e^{iM^{1/2}P^j\cdot X}\phi_j(X)$, where $H(X,P^j) = E$. Assume that a solution to the Schrödinger equation has the asymptotic WKB-form in the hitting plane

$$\Phi(X) = \sum_{\nu=1}^{K}\phi_\nu(X,x)e^{iM^{1/2}\theta_\nu(X)} + \mathcal{O}(M^{-n}),\ \text{for } X_{1_1} = 0 \text{ and for all } n \in \mathbb{N}, \tag{5.1}$$

where $\phi_\nu$ and $\theta_\nu$ are smooth functions, based on a finite sum of WKB-solutions (3.1) (or the caustic ansatz (4.10)).

**Step 3. Use paths to extend the initial Lagrangian manifold.** Use the characteristics paths in Theorem 3.1 for the WKB-functions $\phi_\nu(X_t)$ and $\theta_\nu(X_t)$ (or (4.13) and (4.16) for the caustic case $\theta_\nu(X) = \check{X}\cdot\check{P} - \theta^*(\hat{X},\check{P})$) to locally extend the guessed initial $3N-1$ dimensional Lagrangian manifold, in the hitting plane $X_{1_1} = 0$, to dimension $3N$ outside $X_{1_1} = 0$, by starting a path from each point on the initial Lagrangian manifold; change coordinates in a neighborhood of a caustic and apply the stationary phase method in Step 4 to continue the solutions until the first hitting time $\tau$, for all possible initial $\phi_\nu$ and $\theta_\nu$. Here $\tau$ is the first time the path $(X_t,P_t)$ leaves the set $X_{1_1} < 0$ if it initially went into the set $X_{1_1} > 0$ (i.e. if $\lim_{t\to 0+}\mathrm{sign}(X_{1_1}(t)) = 1$), and similarly the first time it leaves $X_{1_1} > 0$ if $\lim_{t\to 0+}\mathrm{sign}(X_{1_1}(t)) = -1$. For all initial $\phi_\nu$ and $\theta_\nu$ this yields an asymptotic Schrödinger solution

$$\Phi(X) \simeq \sum_\nu \phi_\nu(X,x)e^{iM^{1/2}\theta_\nu(X)} \tag{5.2}$$

for all $X_{1_1} \neq 0$, since by construction the WKB integral is an asymptotic solution

$$(\mathcal{H} - E)\sum_\nu \phi_\nu(X,x)e^{iM^{1/2}\theta_\nu(X)} = \mathcal{O}(M^{-n})$$

in the domain $X_{1_1} \neq 0$ and we assume there exists a stable global solution $\Phi$ (including $X_{1_1} = 0$) to the Schrödinger equation (2.2), for $E$ chosen to be an eigenvalue with a distance bigger than $\mathcal{O}(M^{-n})$ to other eigenvalues. The particular $\phi_\nu$ and $\theta_\nu$ that gives the asymptotic eigensolution satisfies

$$\lim_{X_{1_1}\to 0+}\Phi(X) = \lim_{X_{1_1}\to 0-}\Phi(X). \tag{5.3}$$

We have used the WKB-method to characterize the global solution. In particular the initial Lagrangian manifold, given by $\{\theta_\nu \mid \forall \nu\}$, is by the characteristics mapped to the same manifold at the hitting surface $X_{1_1} = 0$, since otherwise (5.3) and $\lim_{X_{1_1}\to 0+} \theta_\nu(X) = \lim_{X_{1_1}\to 0-} \theta_\nu(X)$ cannot hold. The obtained fixed point initial data $\Phi$ is smooth, since it is a solution of the Schrödinger eigenvalue problem. We assume that also the amplitudes $\phi_\nu$ and the phases $\theta_\nu$ are smooth functions and that there are only finitely many of them. The construction of WKB solutions makes it possible to compare their Lagrangian manifolds by their Hamiltonians. We interpret the molecular dynamics as the formal limit $M \to \infty$ of the WKB system.

**Step 4. Apply Maslov's matching around caustics.** We see that the weight function $G$, in (3.13), based on a single WKB-mode (3.1) blows up at caustics, where $\det(\partial(\check{X})/\partial(\check{P})) = 0$, and that the weight function $G$ in (4.17) for the Fourier integral (4.10) blows up at points where $\det(\partial(\check{P})/\partial(\check{X}))$ vanishes. Therefore, in neighborhoods around caustic points we need to use the representation $\theta^*(\hat{X}, \check{P})$ of the phase based on the Fourier integrals, while around points where $\det(\partial(\check{P})/\partial(\check{X}))$ vanishes we apply the representation $\theta(\hat{X}, \check{X})$ based on the Legendre transform, as pointed out by Maslov in [25] and described in the simplifying setting of the harmonic oscillator in [11].

One way to make a global construction of a WKB solution, in the spirit of [25], is to use the characteristics and a partition of the phase-space as follows, also explained constructively by the numerical algorithm 9.2 in the next section. Start with a Fourier integral representation in a neighborhood $\mathcal{U}$ of a caustic point, which gives a representation of the Schrödinger solution $\Phi$ in $\mathcal{U}$. Then we use the stationary phase expansion, see Section 10, to find an asymptotic approximation $\tilde{\Phi}$ (accurate to any order $n \in \mathbb{N}$) at the boundary points $\check{X}$ of $\mathcal{U}$ as a sum of single WKB-modes with phase functions $\theta_j$

$$\int_{\mathbb{R}^d} \chi(\check{P}) e^{-iM^{1/2}(\check{X}\cdot\check{P} - \theta^*(\hat{X},\check{P}))}\, d\check{P} = \sum_j e^{-iM^{1/2}\theta_j(X)}\phi_j(X) + \mathcal{O}(M^{-n})$$

where each phase function $\theta_j(X) := \check{X} \cdot \check{P}_{X,j} - \theta^*(\hat{X}, \check{P}_{X,j})$ corresponds to a branch of the boundary and the index $j$ corresponds to different solutions $\check{P}_{X,j}$ of the stationary phase equation $\check{X} = \nabla_{\check{P}}\theta^*(\hat{X}, \check{P}_X)$. The single WKB-modes $\phi(x, X)e^{iM^{1/2}\theta(X)}$ are then constructed along the characteristics to be Schrödinger solutions in a domain around the point where $\det(\partial(\check{P})/\partial(\check{X}))$ vanishes, following the construction in Theorem 3.1 using the initial data of $\tilde{\Phi}$ at $\partial\mathcal{U}$. We note that the tiny error of size $\mathcal{O}(M^{-n})$ that we make in the initial data for $\phi$ also yields a tiny perturbation error in $\phi$ of size $\mathcal{O}(M^{-n})$ along the path, due to the assumption of the $\mathcal{O}(1)$ bounded hitting times. A small error we make in the expansion therefore leads to a negligible error in the Schrödinger solution and the corresponding density.

When a characteristic leaves the domain and enters another region around a caustic we again use the stationary phase method at the boundary to give initial data for $(X, P, \phi, G)$. When the characteristic finally returns to the first boundary $\partial\mathcal{U}$, there is a compatibility condition to have a global solution, by having the incoming final phase equal to the initial phase function in $\mathcal{C}^1$. We can think of this as trying to find a co-dimension one surface $I$ in $\mathbb{R}^{3N}$ where the incoming and outgoing phases are equal. First to have one point where they agree is possible if we restrict the possible solutions to a discrete set of energies $E$, i.e., the eigenvalues, and therefore the compatibility condition is called a quantization condition. Then, having one point where the difference of the two phase function is zero, we can combine this with the assumption that the Lagrangian manifold generated by the characteristics path $(X_t, P_t)$ is continuous: the two phases have the same gradient on $I$, since $(X, P) = (X, \nabla_X\theta(X)) = \Big(\big(\hat{X}, \nabla_{\check{P}}\theta^*(\hat{X}, \check{P})\big), \big(\nabla_{\hat{X}}\theta^*(\hat{X}, \check{P}), \check{P}\big)\Big)$ so the phases are $\mathcal{C}^1$. If the Lagrangian manifold would be simple connected, the compatibility condition is always satisfied by the construction of the Lagrangian manifold; here we assume that one parameter is enough to describe the non simple connectedness. In this way we define the $(X, P, \phi, G)$ globally, for the eigenvalue energies $E$. To evaluate observables we use a partition of unity to restrict the observable to a domain with a single representation, either a Fourier integral representation for a caustic or a single WKB-mode when $\det(\partial(\check{P})/\partial(\check{X})) = 0$.

5.1. **Lagrangian manifolds.** This section presents some basic properties of Lagrangian manifolds. Given the $3N-1$ dimensional Lagrangian manifold $L$ on the hitting surface $X_{1_1} = 0$, the solution paths

$$\{(X_t, P_t) \in \mathbb{R}^{6N} \,|\, 0 \le t < \infty,\, \forall (X_0, P_0) \in L\}$$

of the Hamiltonian system

$$\dot{X}_t = \nabla_P H(P_t, X_t)$$
$$\dot{P}_t = -\nabla_X H(P_t, X_t)$$

with a smooth and bounded Hamiltonian $H(P, X)$ generate a $3N$-dimensional manifold called Lagrangian manifold. The fact that the Lagrangian manifold has dimension $3N$ implies that it can locally be described by $(X, P)$ with $P$ as a function of $X$ or with $X$ as a function of $P$ and in general $3N$ coordinates are functions of the other $3N$ coordinates.

The Lagrangian manifold generated by the tube of trajectories is defined by the phase function $\theta(X)$ that plays the role of a generating function of the Lagrangian manifold. Thus we seek a function $\theta : U \subset \mathbb{R}^{3N} \to \mathbb{R}$ such that $P_t = \nabla_X \theta(X_t)$. We show that there exists a potential function $\theta$ by determining an equation that preserves the symmetry for the matrix $Q_t$, defined as $Q^{ij}(X) := \partial_{X^j} P^i(X)$ and $Q_t^{ij} := Q^{ij}(X_t)$. The relations $P_t^i = P^i(X_t)$ and $Q_t^{ij} := \partial_{X^j} P^i(X_t)$ imply

$$\dot{P}_t^i = \frac{d}{dt} P^i(X_t) = \sum_j \dot{X}_t^j \partial_{X^j} P_t^i = \sum_j \dot{X}_t^j Q_t^{ij} = \sum_j \partial_{P^j} H\big(P(X_t), X_t\big) Q_t^{ij} \,,$$

so that

$$\begin{aligned}
\partial_{X^k} \dot{P}_t^i &= \partial_{X^k} \Big( \sum_j \partial_{P^j} H\big(P(X_t), X_t\big) Q_t^{ij} \Big) \\
&= \underbrace{\sum_j \dot{X}_t^j \partial_{X^k} Q_t^{ij}}_{= \sum_j \dot{X}_t^j \partial_{X^k X^j} P_t^i = \sum_j \dot{X}_t^j \partial_{X^j X^k} P_t^i = \dot{Q}_t^{ik}} + \sum_j \partial_{P^j P^l} H\big(P(X_t), X_t\big) \underbrace{\partial_{X^k} P^l}_{= Q^{lk}} Q^{ij} \\
&\quad + \sum_j \partial_{P^j X^k} H\big(P(X_t), X_t\big) Q_t^{ij}
\end{aligned}$$

and

$$\partial_{X^k} \dot{P}_t^i = -\partial_{X^k} \Big( \partial_{X^i} H\big(P(X_t), X_t\big) \Big) = -\partial_{X^i X^k} H\big(P(X_t), X_t\big) \Big) - \sum_j \partial_{X^i P^j} H\big(P(X_t), X_t\big) \underbrace{\partial_{X^k} P^j}_{= Q_t^{jk}}$$

together with the symmetry of $Q_t$ show that

$$\begin{aligned}
\dot{Q}_t^{ik} &= -\partial_{X^i X^k} H(P_t, X_t) - \sum_{j,l} \partial_{P^j P^l} H(P_t, X_t) Q_t^{kl} Q_t^{ij} \\
&\quad - \sum_j \partial_{P^j X^k} H(P_t, X_t) Q_t^{ij} - \sum_j \partial_{P^j X^i} H(P_t, X_t) Q_t^{kj} \,.
\end{aligned} \tag{5.4}$$

Since the Hamiltonian is assumed to be smooth it follows that the right hand side in (5.4) is symmetric and thus the matrix $Q_t$ remains symmetric since it is initially symmetric. Hence there exists a potential function $\theta(X)$ such that $P(X) = \nabla_X \theta(X)$ in simple connected domains where $Q$ is smooth. The function $Q$ may become unbounded due to the term $\partial_{P^j P^l} H\, Q^{kl} Q^{ij}$, even though $H$ has bounded third derivatives. Points $X_t$ at which $|\mathrm{Tr}\,(Q_t)| = \infty$ satisfy, by Liouville's theorem (see Section 3.1.4), $\left|\det \frac{\partial X_0}{\partial X_t}\right| = \infty$ and such points are called *caustic* points.

The same construction of a potential works for the local chart expressed as $X = X(P)$ instead of $P = P(X)$. In fact any new variable $\hat{X}$ (not including both $X^i$ and $P^i$ for any $i$), based on $3N$ of the $6N$ variables $(X, P)$, and the remaining variables $3N$ variables, $\hat{P}$, represent the same Hamiltonian system with the Hamiltonian $\hat{H}(\hat{P}, \hat{X}) := H(P, X)$. The Lagrangian manifold is defined by $\hat{P} = \nabla_{\hat{X}} \hat{\theta}(\hat{X})$ in the local chart of $\hat{P}$-coordinates with the generating (potential) function $\hat{\theta}(\hat{X})$ defined in domains excluding caustics, i.e., where $\det \left|\frac{\partial \hat{X}_0}{\partial \hat{X}_t}\right| < \infty$. Maslov, [25], realized that a Lagrangian manifold can be partitioned, by changing coordinates in the neighborhood of a caustic, into domains where $\hat{P} = \nabla_{\hat{X}} \hat{\theta}(\hat{X})$ is smooth. He used the generating (potential) functions $\hat{\theta}$ to construct asymptotic WKB solutions of Fourier integral type. A sketch of this general situation is depicted in Figure 4. In previous sections we have described construction of

solutions in a simpler case without caustics, i.e., $P_t = \nabla_X \theta(X_t)$ holds everywhere. In this section we described the global construction of WKB solutions in the general case when caustics are present.

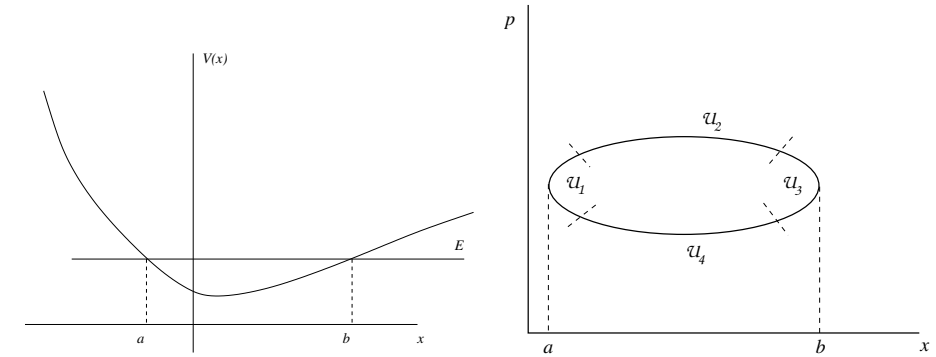


FIGURE 4. The left figure depicts a graph of the molecular dynamics potential $\lambda(X)$ in the case which exhibits caustics at $X = a$ and $X = b$ for a given energy $E$. The right figure shows a general case of the Lagrangian manifold with two caustic points $X = a$ and $b$ and its covering with charts $\mathcal{U}_i$. In the charts $\mathcal{U}_i$, $i = 2, 4$ the manifold is defined by $P = \nabla_X \theta_i(X)$ and the solution to Schrödinger equation is constructed by simple WKB modes. The caustics belong to the charts $\mathcal{U}_i$, $i = 1, 3$ and in this case the manifold is defined by $X = \nabla_X \theta_i(P)$ and the solutions are given by the Fourier integrals.

## 6. Computation of observables

Suppose the goal is to compute a real-valued *observable*

$$\int_{\mathbb{T}^{3N}} \langle \Phi, A\Phi \rangle \, dX$$

for a given bounded linear multiplication operator $A = A(X)$ on $L^2(\mathbb{T}^{3N})$ and a solution $\Phi = \sum_k \phi_k e^{iM^{1/2}\theta_k}$ of (2.2). We have

$$\begin{aligned} \textstyle\int_{\mathbb{T}^{3N}} \langle \Phi, A\Phi \rangle dX &= \textstyle\sum_{k,l} \int_{\mathbb{T}^{3N}} \langle A\phi_k e^{iM^{1/2}\theta_k(X)}, \phi_l e^{iM^{1/2}\theta_l(X)} \rangle \, dX \\ &= \textstyle\sum_{k,l} \int_{\mathbb{T}^{3N}} A e^{iM^{1/2}(\theta_l(X) - \theta_k(X))} \langle \phi_k, \phi_l \rangle \, dX \, . \end{aligned} \tag{6.1}$$

The integrand is oscillatory for $k \neq l$, hence critical points (or near critical points) of the phase difference give the main contribution. The stationary phase method, see [10, 25] and Section 10, shows that these integrals are small, bounded by $\mathcal{O}(M^{-3N/4})$, in the case when the phase difference has non degenerate critical points, or no critical point, and the functions $A\langle \phi_k, \phi_l \rangle$ and $\theta_l$ are sufficiently smooth. A critical point $X_c \in \mathbb{R}^{3N}$ satisfies $\nabla_X \theta_l(X_c) - \nabla_X \theta_k(X_c) = 0$, which means that the two different paths, generated by $\theta_l$ and $\theta_k$, passing through $X = X_c$ also have the same momentum $P$ at this point. That the critical point is degenerate means that the Hessian matrix $\partial_{X^i X^j}(\theta_k - \theta_l)(X_c)$ is singular (or asymptotically singular for $M \to \infty$ as for avoided crossings when the electron eigenvalues have a vanishing spectral gap depending on $M$). Therefore caustics, crossing or avoided crossing electron eigenvalues may generate coupling between the WKB terms. On the other hand, without such coupling the density of a linear combination of WKB terms separates asymptotically to a sum of densities of the individual WKB terms

$$\int_{\mathbb{T}^{3N}} \langle \Phi, A\Phi \rangle dX = \int_{\mathbb{T}^{3N}} \underbrace{A \sum_{k=1}^{\bar{k}} \underbrace{\langle \phi_k, \phi_k \rangle}_{=\rho_k}}_{=\rho} \, dX + \mathcal{O}(M^{-1}) \, , \tag{6.2}$$

in the case of multiple WKB-functions, $\bar{k} > 1$, and

$$\int_{\mathbb{T}^{3N}} \langle \Phi, A\Phi \rangle \, dX = \int_{\mathbb{T}^{3N}} A\langle \phi_1, \phi_1 \rangle \, dX$$

for a single WKB-function and we have seen in (4.24) that the Fourier integral WKB-solutions around caustics yields the same density as a single WKB mode.

In the presence of a caustic, the WKB terms can be asymptotically non orthogonal, since their coefficients and phases typically are not smooth enough to allow the integration by parts to gain powers of $M^{-1/2}$. Non-orthogonal WKB functions tell how the caustic couples the WKB modes.

Regarding the inflow density $\rho_k\big|_I$ there are two situations: either the characteristics return often to the inflow domain or not. If they do not return we have a scattering problem and it is reasonable to define the inflow-density $\rho_k\big|_I$ as an initial condition. If characteristics return, the dynamics can be used to estimate the return-density $\rho_k\big|_I$ as follows: Assume that the following limits exist

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T A(X_t) \, dt = \int_{\mathbb{T}^{3N}} A(X)\rho_k(X) \, dX \tag{6.3}$$

which bypasses the need to find $\rho_k\big|_I$ and the quadrature in the number of characteristics. When there are multiple amplitudes $\phi_k$ in (6.2) we can, away from caustics, obtain the sum of the densities $\sum_k \rho_k$ from the time average split into the time the dynamics spends in each phase $\theta_k$. A way to think about the time average (6.3) is to sample the return points $X_t \in I$ and from these samples construct an *empirical* return-density, converging to $\rho_k\big|_I$ as the number of return iterations tends to infinity. We shall use this perspective to view the eikonal equation (3.8) as a hitting problem on $I$, with hitting times $\tau$ (i.e., return times). The stronger property having $\rho_k$ constant as a function of $(X_0, P_0)$ for $H(X_0, P_0) = E$ is called *ergodicity*, see [30]. We allow the density $\rho_k|_I$ to depend on the initial position $X_0$ and momentum $P_0$ and then our observables are conditional expected values. An example of a hitting surface is the co-dimension one surface where the first component $X_{1_1}$ in $X_1 = (X_{1_1}, X_{1_2}, X_{1_3})$ is equal to its initial value $X_{1_1}(0)$. The dynamics does not always have such a hitting surface: for instance if all particles are close initially and then are scattered away from each other, as in an explosion, no co-dimension one hitting surface exists.

## 7. Molecular dynamics approximation of Schrödinger observables

A numerical computation of an approximation to $\sum_k \int_{\mathbb{T}^{3N}} \langle \phi_k, A\phi_k \rangle \, dX$ has the main ingredients:

(1) to approximate the exact characteristics by molecular dynamics characteristics (3.10),
(2) to discretize the molecular dynamics equations, and
(3a) if $\rho\big|_I$ is an inflow-density, to introduce quadrature in the number of characteristics, or
(3b) if $\rho\big|_I$ is a return-density, to replace the ensemble average by a time average using the property (6.3).

This section presents a derivation of the approximation error in the step (1) in the case of a return density and comments on the time-discretization of step (2) treated in Section 7.2. The third and fourth discretization steps, which are not described here, are studied, for instance, in [8, 7, 21].

### 7.1. The Born-Oppenheimer approximation error.

This section states our main result of molecular dynamics approximating Schrödinger observables based on hitting time and spectral gap assumptions. We formulate it using the assumption of the Born-Oppenheimer property

$$\|\psi_t - \Psi_{\mathrm{BO}}(X_t)\|_{L^2(dx)} = \mathcal{O}(M^{-1/2}), \quad \text{uniformly in } t. \tag{7.1}$$

This assumption is then proved in Lemma 8.2 based on a setting with a spectral gap.

*The spectral gap condition.* The electron eigenvalues $\{\lambda_k\}$ satisfy, for some positive $c$, the spectral gap condition

$$\inf_{k\neq 0,\ Y\in D} |\lambda_k(Y) - \lambda_0(Y)| > c, \tag{7.2}$$

where $D := \{X_{\mathrm{S}}(t) \,|\, t \geq 0\} \cup \{X_{\mathrm{BO}}(t) \,|\, t \geq 0\}$ is the set of all nuclei positions obtained from the Schrödinger characteristics $X = X_{\mathrm{S}}$ in Theorem 3.1 and from the Born-Oppenheimer dynamics $X = X_{\mathrm{BO}}$ in (3.16), for all considered initial data.

*The hitting time assumption.* In the case of quasiperiodic dynamics, the points $(X_{t_k}, P_{t_k})$ in a codimension one hitting plane, e.g. $X_1^1 \equiv X_{1_1} = 0$, accumulate on a set of dimension $3N-1$, see Figure 5, while for ergodic dynamics the hitting points are distributed in the phase-space set $\{(X,P) \in \mathbb{R}^{6N} \,|\, H(X,P) = E \;\&\; X_{1_1} = 0\}$ of dimension $6N-2$ with positive density, see Figure 6.

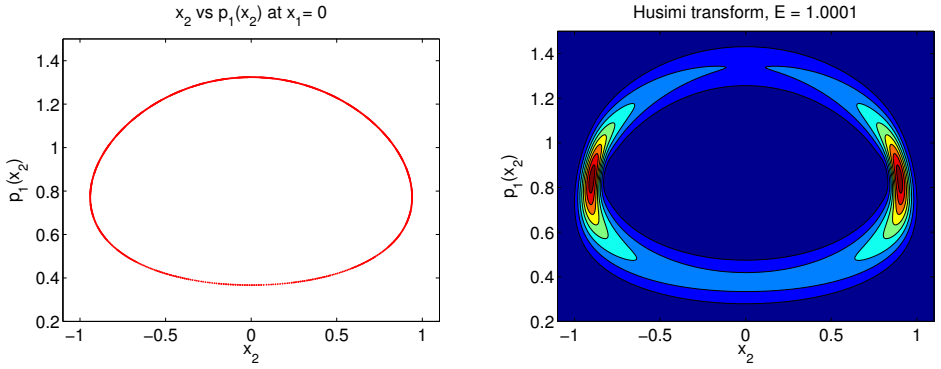


FIGURE 5. Left plot: The hitting points $(X_2, P_1)$ in the plane $X_1 = 0$ that accumulate on a curve of dimension one, for the Hamiltonian $H = |P|^2/2 + |X_1|^2/2 + |X_2|^2/2^{1/2} + 0.3\sin(X_1X_2) = 1.0001$. Right plot: The Husimi transform, $M^{1/2}|\int_{\mathbb{R}} \Phi(0-y, X_2)e^{-iM^{1/2}yP_1}e^{-M^{1/2}|y|^2/2}dy|^2$, of the Schrödinger solution, $\Phi(X)$, as a function of $X_2$ and $P_1$ in the plane $X_1 = 0$, with the potential $V(X) = |X_1|^2/2 + |X_2|^2/2^{1/2} + 0.3\sin(X_1X_2)$ and the eigenvalue $E = 1.0001$ for mass $M = 8000$. The Husimi transform measures the density in $X_2$ and $P_1$ and equals the absolute value squared of the FBI transform, which here is integrated in the $P_2$ direction.

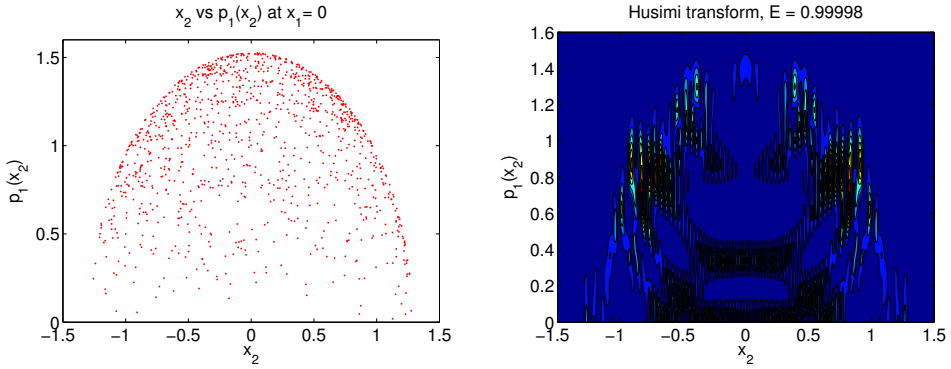


FIGURE 6. Left: The hitting points $(X_2, P_1)$ in the plane $X_1 = 0$ that become dens i set of dimension two, for the Hamiltonian $H = |P|^2/2 + |X_1|^2/2 + |X_2|^2/2^{1/2} + 2\sin(X_1X_2) = 1$. Right: The Husimi transform of the Schrödinger solution in the plane $X_1 = 0$ for the same potential and eigenvalue $E = 0.99998$.

In the quasiperiodic setting it is therefore reasonable to assume that the hitting time (i.e. the return time) is finite, since only a uniformly bounded set of phases in the hitting plane is visited.

**Theorem 7.1.** *Assume that there is an asymptotic solution to the Schrödinger equation* (2.2) *that can be written as a finite sum of WKB-modes* (5.1) *(as for caustics* (4.10)*) and that all phase functions* $\theta_{\mathrm{S}}$ *and* $\theta_{\mathrm{BO}}$ *are smooth solutions to the Schrödinger eikonal equation (written as* (3.8) *or* (4.13)*) and the molecular dynamics eikonal equation* (3.17)*, respectively, satisfying the Born-Oppenheimer property* (7.1) *and the limit* (6.3)*, with uniformly bounded hitting times* $\tau$*, in* (8.12), (8.16)*, and* (8.20)*, then the zero-order*

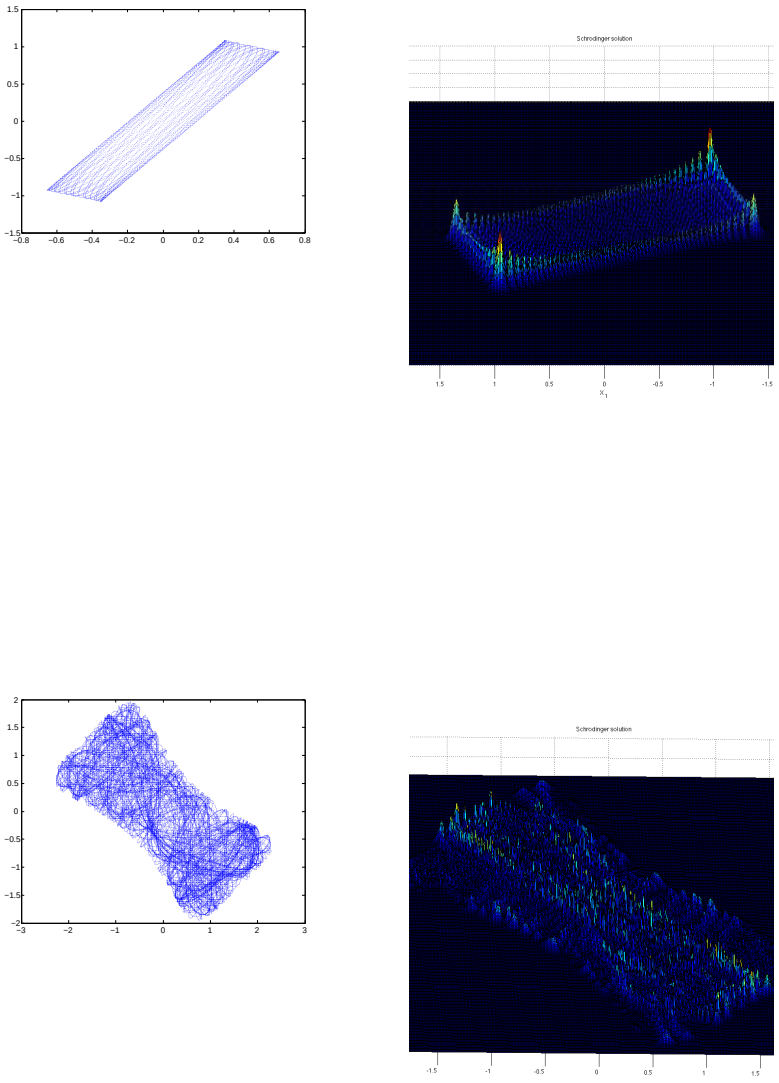


FIGURE 7. Left: the path $(X_1(t), X_2(t))$ for $\beta = 0.3$ and $\beta = 2$. Right: the Schrödinger density $|\Phi(X)|^2$ for $\beta = 0.3$ and $\beta = 2$.

*Born-Oppenheimer dynamics* (3.16)*, with initial data generated by the construction in Steps 1-4 in Section 5, approximates time-independent Schrödinger observables, with error bounded by* $\mathcal{O}(M^{-1+\delta})$*, i.e.*

$$\int_{\mathbb{T}^{3N}} g(X)\rho_{\mathrm{BO}}(X)\,dX = \int_{\mathbb{T}^{3N}} g(X)\rho_{\mathrm{S}}(X)\,dX + \mathcal{O}(M^{-1+\delta})\,, \quad \textit{for any } \delta > 0 \textit{ and } g \in \mathcal{C}^3(\mathbb{R}^{3N}). \tag{7.3}$$

The proof is given in Section 8.

7.2. **Why do symplectic numerical simulations of molecular dynamics work?** The derivation of the approximation error for the Born-Oppenheimer dynamics, in Theorem 7.1, also allows to study perturbed systems. For instance, the perturbed Born-Oppenheimer dynamics

$$\begin{aligned}\dot{X}_t &= P_t + \nabla_P H^\epsilon(X_t, P_t)\\ \dot{P}_t &= -\nabla_X \lambda_0(X_t) - \nabla_X H^\epsilon(X_t, P_t)\,,\end{aligned}$$

generated from a perturbed Hamiltonian $H_{\rm BO}(X,P) + H^\epsilon(X,P) = E$, with the perturbation satisfying

$$\|H^\epsilon\|_{L^\infty} \leq \epsilon \quad \text{for some } \epsilon \in (0,\infty) \tag{7.4}$$

yields through (8.14) and (8.21) an additional error term $\mathcal{O}(\epsilon)$ to the approximation of observables in (7.3). So called symplectic numerical methods are precisely those that can be written as perturbed Hamiltonian systems, see [31], and consequently we have a method to precisely analyze their numerical error by combining an explicit construction of $H^\epsilon$ with the stability condition (7.4) to obtain $\mathcal{O}\big((M^{-1}+\epsilon)^{1-\delta}\big)$ accurate approximations, provided the corresponding phase function has bounded second difference quotients. The popular Störmer-Verlet method is symplectic and the positions $X$ coincides with those of the symplectic Euler method, for which $H^\epsilon$ is explicitly constructed in [31] with $\epsilon$ proportional to the time step. The construction in [31] is not using the modified equation and formal asymptotics, instead a piecewise linear extension of the solution generates $H^\epsilon$.

## 8. Analysis of the molecular dynamics approximation

Before we proceed with the analysis of the approximation error we motivate our results by a significantly simpler case of a system *without electrons*. We use the densities (3.18) and (3.19) and we show heuristically how the characteristics can be used to estimate the difference $\rho_{\rm S} - \rho_{\rm BO}$, leading to $\mathcal{O}(M^{-1})$ accurate Born-Oppenheimer approximations of Schrödinger observables

$$\int g(X) \underbrace{\rho_{\rm S}(X)}_{\langle\Phi,\Phi\rangle}\, dX = \int g(X)\rho_{BO}(X)\, dX + \mathcal{O}(M^{-1})\,.$$

In the special case of no electrons, the dynamics of $X$ does not depend on $\psi$ and therefore $X_{\rm BO} = X_{\rm S} = X$ and consequently $G_{\rm BO} = G_{\rm S}$. The difference $\psi_{\rm S} - \psi_{\rm BO}$ can be understood from iterative approximations of (3.12)

$$\frac{i}{M^{1/2}}\dot{\psi}_{k+1} - (\mathcal{V} - V_0)\psi_{k+1} = \frac{1}{2M} G\Delta_X(G^{-1}\psi_k) \tag{8.1}$$

with $\psi_0 = 0$. Then $\psi_{\rm BO} = \psi_1$ is the Born-Oppenheimer approximation and formally we have the iterations approaching the full Schrödinger solution as $k \to \infty$.

In the special case of no electrons, there holds $\mathcal{V} = V_0$, thus the transport equation $i\dot{\psi}_1 = 0$ has constant solutions. We let $\psi_1 = 1$ and then $\psi_2 - \psi_1$ is imaginary with its absolute value bounded by $\mathcal{O}(M^{-1/2})$. We write the iterations of $\psi_k$ by integrating (8.1) as the linear mapping

$$\psi_{k+1} = 1 + iM^{-1/2}\hat{\mathcal{S}}(\psi_k) = \sum_{l=0}^{k} i^l M^{-l/2}\hat{\mathcal{S}}^l(\psi_1)\,,$$

which formally shows that

$$|\psi_{\rm S}|^2 = |\psi_1|^2 + 2{\rm Re}\,\langle \psi_{\rm S} - \psi_1, \psi_1\rangle + |\psi_{\rm S} - \psi_1|^2 = 1 + \mathcal{O}(M^{-1})\,.$$

Consequently this special Born-Oppenheimer density satisfies

$$\rho_{\rm BO} = \underbrace{G_{\rm S}^{-2}\langle \psi_{\rm S}, \psi_{\rm S}\rangle}_{=\rho_{\rm S}} + \mathcal{O}(M^{-1})\,, \tag{8.2}$$

since $G_{\rm BO} = G_{\rm S}$ and $X$ do not depend on $\psi$.

In the general case with electrons and a spectral gap, we show in Lemma 8.2 that there is a solution $\psi_{\rm S}$ satisfying

$$\|\psi_{\rm S} - \Psi_{\rm BO}\|_{L^2(dx)} = \mathcal{O}(M^{-1/2})\,, \tag{8.3}$$

for the electron eigenfunction $\Psi_{\rm BO}$, satisfying

$$\mathcal{V}(\cdot, X)\Psi_{\rm BO}(\cdot, X) = \lambda_0(X)\Psi_{\rm BO}(\cdot, X)$$

and the eigenvalue $\lambda_0(X) \in \mathbb{R}$ with a (fixed) nuclei position $X$. Then the state $\psi_1$ equal to a constant, in the case of no electrons, corresponds to the electron eigenfunction $\Psi_{\rm BO}$ in the case with electrons present. In the general case the $X$ dynamics for the Schrödinger and the Born-Oppenheimer dynamics are not the same, but we will show that (8.3) implies that the Hamiltonians $H_{\rm S}$ and $H_{\rm BO}$ are $\mathcal{O}(M^{-1})$ close. Using stability of Hamilton-Jacobi equations, the phase functions $\theta_{\rm S}$ and $\theta_{\rm BO}$ are then also close in the maximum norm, which, combined with an assumption of smooth phase functions, show that $|G_{\rm S} - G_{\rm BO}| = \mathcal{O}(M^{-1+\delta})$ for any $\delta > 0$. Lemma 8.2 also shows that $|\langle\psi_{\rm S}, \psi_{\rm S}\rangle - 1| = \mathcal{O}(M^{-1})$ and consequently the density bound $|\rho_{\rm S} - \rho_{\rm BO}| = \mathcal{O}(M^{-1+\delta})$ holds. To obtain the estimate (8.3) the important new property, compared to no electrons, is to use oscillatory cancellation in directions orthogonal to $\Psi_{\rm BO}$.

8.1. **Continuation of the construction of the solution operator.** This section continues the construction of the solution operator started in Section 3.4. Assume for a moment that $\tilde{\mathcal{V}}$ is independent of $\tau$. Then the solution to (3.23) can be written as a linear combination of the two exponentials

$$ae^{i\tau\mathcal{A}_+} + be^{i\tau\mathcal{A}_-}$$

where the two characteristic roots are the operators

$$\mathcal{A}_\pm = (P_1^1)^2\left(-1 \pm (1 - 2(P_1^1)^{-2}\tilde{\mathcal{V}})^{1/2}\right)\,.$$

We see that $e^{i\tau\mathcal{A}_-}$ is a highly oscillatory solution on the fast $\tau$-scale with

$$\lim_{P_1^1\to\infty} \frac{1}{(P_1^1)^2}\mathcal{A}_- = -2{\rm Id}\,,$$

while

$$\lim_{P_1^1\to\infty} \mathcal{A}_+ = -\tilde{\mathcal{V}}\,, \tag{8.4}$$

in distribution sense. Therefore we chose initial data

$$i\dot{\psi}|_{\tau=0} = -\mathcal{A}_+\psi|_{\tau=0} \tag{8.5}$$

to have $b = 0$, which eliminates the fast scale, and the limit $P_1^1 \to \infty$ determines the solution by the Schrödinger equation

$$i\dot{\psi} = \tilde{\mathcal{V}}\psi\,.$$

The next section presents an analogous construction for the slowly, in $\tau$, varying operator $\tilde{\mathcal{V}}$.

8.1.1. *Spectral decomposition.* Write (3.23) as the first order system

$$\begin{aligned} i\dot{\psi} &= \pi \\ i\dot{\pi} &= -2(P_1^1)^2(\tilde{\mathcal{V}}\psi - \pi)\,, \end{aligned}$$

which for $\bar{\psi} := (\psi, \pi)$ takes the form

$$\dot{\bar{\psi}} = i\mathcal{B}\bar{\psi}\,, \quad \mathcal{B} := \begin{pmatrix} 0 & -1 \\ 2(P_1^1)^2\tilde{\mathcal{V}} & -2(P_1^1)^2 \end{pmatrix}\,,$$

where the eigenvalues $\Lambda_\pm$ , right eigenvectors $\mathcal{Q}_\pm$ and left eigenvectors $\mathcal{Q}_\pm^{-1}$ of the real "matrix" operator $\mathcal{B}$ are

$$\Lambda_\pm := (P_1^1)^2 \left( -\mathrm{Id} \pm \left( \mathrm{Id} - 2(P_1^1)^{-2}\tilde{\mathcal{V}} \right)^{1/2} \right),$$
$$\mathcal{Q}_+ := \begin{pmatrix} \mathrm{Id} \\ -\Lambda_+ \end{pmatrix}, \quad \mathcal{Q}_- := \begin{pmatrix} -\Lambda_-^{-1} \\ \mathrm{Id} \end{pmatrix},$$
$$\mathcal{Q}_+^{-1} := (\mathrm{Id} - \Lambda_+\Lambda_-^{-1})^{-1} \begin{pmatrix} \mathrm{Id} \\ \Lambda_-^{-1} \end{pmatrix}, \quad \mathcal{Q}_-^{-1} := (\mathrm{Id} - \Lambda_+(\Lambda_-)^{-1})^{-1} \begin{pmatrix} \Lambda_+ \\ \mathrm{Id} \end{pmatrix}.$$

We see that $\lim_{P_1^1\to\infty} \Lambda_+ = -\tilde{\mathcal{V}}$ and $\lim_{P_1^1\to\infty} (P_1^1)^{-2}\Lambda_- = -2\mathrm{Id}$. The important property here is that the left eigenvector limit $\lim_{P_1^1\to\infty} \mathcal{Q}_+^{-1} = (\mathrm{Id}, 0)$ is constant, independent of $\tau$, which implies that the $\mathcal{Q}_+$ component $\mathcal{Q}_+^{-1}\bar{\psi} = \psi$ decouples. We obtain in the limit $P_1^1 \to \infty$ the time-dependent Schrödinger equation

$$\begin{aligned} i\dot{\psi}(\tau) = i\frac{d}{d\tau}(\mathcal{Q}_+^{-1}\bar{\psi}_\tau) = i\mathcal{Q}_+^{-1}\frac{d}{d\tau}\bar{\psi}_\tau = & -\mathcal{Q}_+^{-1}\mathcal{B}_\tau\bar{\psi}_\tau \\ = & -\Lambda_+(\tau)\mathcal{Q}_+^{-1}\bar{\psi}_\tau = -\Lambda_+(\tau)\psi(\tau) = \tilde{\mathcal{V}}_\tau\psi(\tau), \end{aligned}$$

where the operator $\tilde{\mathcal{V}}_\tau$ depends on $\tau$ and $(x, X_0)$, and we define the solution operator $\mathcal{S}$

$$\psi(\tau) = \mathcal{S}_{\tau,0}\psi(0). \tag{8.6}$$

As in (8.5) we can view this as choosing special initial data for $\psi(0)$. From now on we only consider such data.

The operator $\tilde{\mathcal{V}}$ can be symmetrized

$$\bar{\mathcal{V}}_\tau := G_\tau^{-1}\tilde{\mathcal{V}}_\tau G_\tau = (\mathcal{V} - V_0)_\tau - \frac{1}{2M}\sum_j \Delta_{X_*^j}, \tag{8.7}$$

with real eigenvalues $\{\check{\lambda}_m\}$ and orthonormal eigenvectors $\{\zeta^m\}$ in $L^2(dx\, dX_*)$, satisfying

$$\bar{\mathcal{V}}_\tau\zeta^m(\tau) = \check{\lambda}_m(\tau)\zeta^m(\tau).$$

Therefore $\tilde{\mathcal{V}}_\tau$ has the same eigenvalues and the eigenvectors $\bar{\zeta}^m := G_\tau\zeta^m$, which establishes the spectral representation

$$\tilde{\mathcal{V}}_\tau\psi(\cdot,\tau,\cdot) = \sum_m \check{\lambda}_m(\tau)\int_{\mathbb{T}^{3N-1}} \langle\psi(\cdot,\tau,\cdot), \bar{\zeta}^m\rangle G_\tau^{-2} dX_*\, \bar{\zeta}^m(\tau). \tag{8.8}$$

We note that the weight $G^{-2}$ on the co-dimension one surface $\mathbb{T}^{3N-1}$ appears precisely because the operator $\tilde{\mathcal{V}}$ is symmetrized by $G^{-2}$ and the weight $G^{-2}$ corresponds to the Eulerian-Lagrangian change of coordinates (3.13)

$$\int_{\mathbb{T}^{3N-1}} \langle\psi, \bar{\zeta}^m\rangle G_\tau^{-2}\, dX_* = \int_{\mathbb{T}^{3N-1}} \langle\psi, \bar{\zeta}^m\rangle\, dX_0. \tag{8.9}$$

The existence of the orthonormal set of eigenvectors and real eigenvalues makes the operator $\tilde{\mathcal{V}}$ self-adjoint in the Lagrangian coordinates and hence the solution operator $\mathcal{S}$ becomes unitary in the Lagrangian coordinates.

8.2. **Stability from perturbed Hamiltonians.** In this section we derive error estimates of the weight functions $G$ when the corresponding Hamiltonian system is perturbed. To derive the stability estimate we consider the Hamilton-Jacobi equation

$$H(\nabla_X\theta(X), X) = 0 \tag{8.10}$$

in an optimal control perspective with the corresponding Hamiltonian system

$$\dot{X}_t = \nabla_P H(P_t, X_t)$$
$$\dot{P}_t = -\nabla_X H(P_t, X_t).$$

We define the "value" function

$$\theta(X_0) = \theta(X_t) - \int_0^t h(P_s, X_s)\, ds,$$

where the "cost" function defined by

$$h(P,X) := P \cdot \nabla_P H(P,X) - H(P,X)$$

satisfies the Pontryagin principle (related to the Legendre transform)

$$H(P,X) = \sup_Q \left( P \cdot \nabla_Q H(Q,X) - h(Q,X) \right). \tag{8.11}$$

Let $\theta\Big|_I$ be defined by the hitting problem

$$\theta(X_0) = \theta(X_\tau) - \int_0^\tau h(P_s, X_s)\, ds$$

using the hitting time $\tau$ on the return surface $I$

$$\tau := \inf\{t \,|\, X_0 \in I,\, X_t \in I \,\&\, t > 0\}. \tag{8.12}$$

For a perturbed Hamiltonian $\tilde{H}$ and its dynamics $(\tilde{X}_t, \tilde{P}_t)$ we define analogously the value function $\tilde{\theta}$ and the cost function $\tilde{h}$.

8.2.1. *Estimates of the phase functions.* We can think of the difference $\theta - \tilde{\theta}$ as composed by a perturbation of the boundary data (on the return surface $I$) and perturbations of the Hamiltonians. The difference of the value functions due to the perturbed Hamiltonian satisfies the stability estimate

$$\begin{aligned} \theta(X_0) - \tilde{\theta}(X_0) &\geq \theta(\tilde{X}_{\tilde{\tau}}) - \tilde{\theta}(\tilde{X}_{\tilde{\tau}}) + \int_0^{\tilde{\tau}} (H - \tilde{H}) \left( \nabla_X \theta(\tilde{X}_t), \tilde{X}_t \right) dt \\ \theta(X_0) - \tilde{\theta}(X_0) &\leq \theta(X_\tau) - \tilde{\theta}(X_\tau) + \int_0^{\tau} (H - \tilde{H}) \left( \nabla_X \tilde{\theta}(X_t), X_t \right) dt \end{aligned} \tag{8.13}$$

with a difference of the Hamiltonians evaluated along the same solution path. This result follows by differentiating the value function along a path and using the Hamilton-Jacobi equations, see Remark 8.1 and [9].

The global construction of a Hamilton-Jacobi equation in Section 5 shows that the Hamiltonian $H(X,P)$ remains the same while the parametrization $(X,P) = (\hat{X}, \check{X}, \hat{P}, \check{P})$ of the Lagrangian manifold may vary in different domains - e.g. $P = P(X) = \nabla_X \theta(X)$ or $(\hat{P}, \check{X}) = \left(\nabla_{\hat{X}} \theta^*(\hat{X}, \check{P}), \nabla_{\check{P}} \theta^*(\hat{X}, \check{P})\right)$- due to the effect of caustics. At the interface to such domains the phase function satisfies a boundary condition, which determines the phase function with an asymptotically vanishing error of size $\mathcal{O}(M^{-n})$ for any natural number $n$. This small error in the initial data of the phase function is negligible compared to other sources of error.

We assume that

$$\sup_{(P,X) \in L(H) \cup L(\tilde{H})} |(H - \tilde{H})(P,X)| = \mathcal{O}(M^{-1}), \tag{8.14}$$

which is verified in (8.15) for Schrödinger and Born-Oppenheimer Hamiltonians; here $L(H)$ is the Lagrangian manifold for the Hamiltonian $H$. We consider the case when the return time is uniformly bounded, related to quasiperiodic dynamics.

The Hamiltonians we use are

$$\begin{aligned} H_{\mathrm{S}} &= \frac{|P|^2}{2} + \frac{\langle \psi(X), \mathcal{V}(X)\psi(X) \rangle}{\langle \psi(X), \psi(X) \rangle} - E, \\ H_{\mathrm{BO}} &= \frac{|P|^2}{2} + \lambda_0(X) - E, \end{aligned}$$

based on the cost functions

$$\begin{aligned} h_{\mathrm{S}} &= E + \frac{|P|^2}{2} - \frac{\langle \psi(X), \mathcal{V}(X)\psi(X) \rangle}{\langle \psi(X), \psi(X) \rangle}, \\ h_{\mathrm{BO}} &= E + \frac{|P|^2}{2} - \lambda_0(X). \end{aligned}$$

For the Born-Oppenheimer case the electron wave function is the eigenstate $\Psi_{\mathrm{BO}}$. The Born-Oppenheimer approximation (7.1), proved in Lemma 8.2, implies that

$$\|H_S - H_{BO}\|_{L^\infty} = \mathcal{O}(M^{-1}), \tag{8.15}$$

which verifies (8.14).

*Uniformly bounded hitting times.* We choose the $3N-1$ dimensional hitting set as

$$I := \{X \in \mathbb{T}^{3N} \,|\, \theta(X) = \tilde{\theta}(X)\} \tag{8.16}$$

on which the two phases coincide. If $\theta(X)$ is a solution also $\theta(X)+C$ is a solution to the Hamilton-Jacobi equation (8.10), for any constant $C$; therefore we can choose $\theta$ such that $\theta(X_0) = \tilde{\theta}(X_0)$ for any $X_0$. Now assume that $I$ forms a codimension one set in $\mathbb{T}^{3N}$ and that the maximal hitting time $\tau$ for characteristics starting on $I$ is bounded; the fact that $I$ is a codimension one set holds, for instance, locally if $|\nabla_X(\theta-\tilde{\theta})|$ is nonzero. In fact, it is sufficient to assume that there exists a function $\gamma : \mathbb{T}^{3N} \to \mathbb{R}$, satisfying $\gamma = \mathcal{O}(M^{-1})$, such that the set $I := \{X \in \mathbb{T}^{3N} \,|\, \theta(X) - \tilde{\theta}(X) = \gamma(X)\}$ is a codimension one set with bounded hitting times. We also define the $3N-1$ dimensional hitting sets in phase space

$$I_\theta := \{(X,P) \mid X \in I,\ P = \nabla\theta(X)\},$$
$$I_{\tilde{\theta}} := \{(X,P) \mid X \in I,\ P = \nabla\tilde{\theta}(X)\}.$$

Then the representation (8.13), for any time $t$ replacing $\tau$ and $\tilde{\tau}$, together with the stability of the Hamiltonians (8.14) and the initial data $(\theta-\tilde{\theta})|_I = 0$ obtained from (8.16) imply that

$$\|\theta - \tilde{\theta}\|_{L^\infty} = \mathcal{O}(M^{-1})\,, \tag{8.17}$$

provided the maximal hitting time $\tau$ is bounded: we assume the hitting time is uniformly bounded in the case when the Lagrangian manifold $L$ for the Schrödinger Hamiltonian system in Section 5 has a smooth limit, as $M$ tends to infinity, and the hitting points on a codimension one plane in phase space accumulate on a $3N-1$ dimensional subset of phase space; this case is related to quasiperiodic molecular dynamics systems.

8.2.2. *Estimates of the densities.* To estimate the density, we will use the characteristic paths. When the value functions $\theta$ and $\tilde{\theta}$ are smoothly differentiable in $X$, with derivatives bounded uniformly in $M$, the stability estimate (8.13) implies that also the difference of the second derivatives has the bound

$$\|\Delta_X\theta - \Delta_X\tilde{\theta}\|_{L^\infty} = \mathcal{O}(M^{-1+\delta})\,, \quad \text{for any } \delta > 0. \tag{8.18}$$

Our goal is to analyze the density function $\rho = |G|^{-2}\langle\psi,\psi\rangle$ with $G$ defined in (3.11). The Born-Oppenheimer approximation (7.1) yields $\langle\psi,\psi\rangle = 1 + \mathcal{O}(M^{-1})$ thus it remains to estimate the weight function $|G|^{-2}$. This weight function satisfies the Hamilton-Jacobi equation

$$H_G(\nabla_X \log|G|^{-2}, X) := \nabla_X\theta(X)\cdot\nabla_X \log|G|^{-2} + \Delta_X\theta(X) = 0\,. \tag{8.19}$$

The stability of Hamilton-Jacobi equations can then be applied to (8.19), as in (8.13), using now the hitting set

$$I := \{X \in \mathbb{T}^{3N} \;\big|\; \log|G(X)|^{-2} = \log|\tilde{G}(X)|^{-2}\} \tag{8.20}$$

and the assumption of bounded hitting times $\tau$ in the hitting problem, and we obtain

$$\|\log|G|^{-2} - \log|\tilde{G}|^{-2}\|_{L^\infty} \le C\|H_G - H_{\tilde{G}}\|_{L^\infty} = \mathcal{O}(M^{-1+\delta})\,. \tag{8.21}$$

In this sense we will use that an $\mathcal{O}(M^{-1})$ perturbation of the Hamiltonian yields an error estimate of almost the same order for the difference of the corresponding densities $\rho - \tilde{\rho}$.

**Remark 8.1.** This remark derives the stability estimate (8.13). The definitions of the value functions imply

$$
\begin{aligned}
&\underbrace{\tilde\theta(\tilde X_{\tilde\tau}) - \int_0^{\tilde\tau} \tilde h(\tilde P_t, \tilde X_t)\,dt}_{\tilde\theta(\tilde X_0)} - \underbrace{\Big(\theta(X_\tau) - \int_0^{\tau} h(P_t, X_t)\,dt\Big)}_{\theta(X_0)} \\
&= -\int_0^{\tilde\tau} \tilde h(\tilde P_t, \tilde X_t)\,dt + \theta(\tilde X_{\tilde\tau}) - \underbrace{\theta(X_0)}_{\theta(\tilde X_0)} + \tilde\theta(\tilde X_{\tilde\tau}) - \theta(\tilde X_{\tilde\tau}) \\
&= -\int_0^{\tilde\tau} \tilde h(\tilde P_t, \tilde X_t)\,dt + \int_0^{\tilde\tau} d\theta(\tilde X_t) + \tilde\theta(\tilde X_{\tilde\tau}) - \theta(\tilde X_{\tilde\tau}) \\
&= \int_0^{\tilde\tau} \underbrace{-\tilde h(\tilde P_t, \tilde X_t) + \nabla_X \theta(\tilde X_t)\cdot \nabla_P \tilde H(\tilde P_t, \tilde X_t)}_{\le \tilde H(\nabla_X\theta(\tilde X_t), \tilde X_t)}\,dt + \tilde\theta(\tilde X_{\tilde\tau}) - \theta(\tilde X_{\tilde\tau}) \\
&\le \int_0^{\tilde\tau} (\tilde H - H)\left(\nabla_X\theta(\tilde X_t), \tilde X_t\right)\,dt + \tilde\theta(\tilde X_{\tilde\tau}) - \theta(\tilde X_{\tilde\tau})\,,
\end{aligned}
\tag{8.22}
$$

where the Pontryagin principle (8.11) yields the inequality and we use the Hamilton-Jacobi equation

$$
H(\nabla_X\theta(\tilde X_t), \tilde X_t) = 0\,.
$$

To establish the lower bound we replace $\theta$ along with $\tilde X_t$ by $\tilde\theta$ and $X_t$ and repeat the derivation above.

8.3. **The Born-Oppenheimer approximation.** The purpose of this section is to present a case when the Born-Oppenheimer approximation holds in the sense that $\|\psi - \Psi_{\rm BO}\|_{L^2(dx)}$ is small.

We know from Section 8.1.1 that the solution $\psi_t = \mathcal{S}_{t,0}\psi_0$ is bounded in $L^2(dx\,dX)$, since $\mathcal{S}$ is unitary in the Lagrangian coordinates. This unitary $\mathcal{S}$ implies that the integral in the Lagrangian coordinates $\int_{\mathbb{T}^{3N-1}} \langle\psi_t, \psi_t\rangle\,dX_0$ is constant in time. We consider the co-dimension one set

$$
I_\psi := \{X \in \mathbb{R}^{3N} \,|\, \langle\psi(X), \psi(X)\rangle = \int_{\mathbb{T}^{3N-1}} \langle\psi(t, X_0), \psi(t, X_0)\rangle\,dX_0 / \int_{\mathbb{T}^{3N-1}} dX_0\}\,,
$$

where the point values of $\langle\psi(X), \psi(X)\rangle$ coincides with its $L^2$ average. We choose a time $t$ such that $X_t \in I_\psi$ and assume that the time $\tau^*$ it takes to hit $I_\psi$ the next time is bounded, i.e.,

$$
\tau^* := \inf\{\tau \,|\, X_t \in I_\psi,\ \tau > 0\ \&\ X_{t+\tau} \in I_\psi\} = \mathcal{O}(1)\,.
$$

We also assume that all functions of $X$ are smooth.

**Lemma 8.2.** *Assume that $i\dot\psi = M^{1/2}\tilde{\mathcal{V}}\psi$ holds, then there exists initial data for $\psi$ such that the $L^2(dx)$ orthogonal decomposition $\psi = \bar\psi_0 \oplus \psi_0^\perp$, where $\bar\psi_0 = \alpha\Psi_{\rm BO}$ for some $\alpha \in \mathbb{C}$ satisfies*

$$
\begin{aligned}
\frac{\|\psi_0^\perp(t)\|_{L^2(dx)}}{\|\bar\psi_0(t)\|_{L^2(dx)}} &= \mathcal{O}(M^{-1/2}) \\
|\langle\psi_t, \psi_t\rangle - 1| &= \mathcal{O}(M^{-1}) \\
\|\psi_t - \Psi_{\rm BO}(X_t)\|_{L^2(dx)} &= \mathcal{O}(M^{-1/2})
\end{aligned}
\tag{8.23}
$$

*uniformly in time $t$, provided the spectral gap condition* (7.2) *holds, the smoothness estimate* (8.30) *is satisfied and the hitting time $\tau^*$ is bounded.*

*Proof.* We consider the decomposition $\psi = \bar\psi_0 \oplus \psi_0^\perp$, where $\bar\psi_0(\tau)$ is an eigenfunction of $\mathcal{V}(X_\tau)$ in $L^2(dx)$, satisfying $\mathcal{V}(X_\tau)\bar\psi_0(\tau) = \lambda_0(\tau)\bar\psi_0(\tau)$ for the eigenvalue $\lambda_0(\tau) \in \mathbb{R}$. This *ansatz* is motivated by the zero residual

$$
\mathcal{R}\psi := \dot\psi + iM^{1/2}\tilde{\mathcal{V}}\psi = 0 \tag{8.24}
$$

and the small residual for the eigenfunction

$$
\begin{aligned}
\langle\Pi(\dot{\bar\psi}_0), \bar\psi_0\rangle &= 0 \\
M^{1/2}\tilde{\mathcal{V}}\bar\psi_0 &= \mathcal{O}(M^{-1/2})\,,
\end{aligned}
$$

where

$$w(X) = \langle \Psi_{\mathrm{BO}}(X), w(X)\rangle \Psi_{\mathrm{BO}}(X) \oplus \Pi w(X) \tag{8.25}$$

denotes the orthogonal decomposition in the eigenfunction direction $\Psi_{\mathrm{BO}}$ and its orthogonal complement in $L^2(dx)$. We consider first the linear operator $\mathcal{R}$ in (8.24) with a given function $V_0$ and then we use a contraction setting to show that $V_0 = \langle \psi, \mathcal{V}\psi\rangle / \langle \psi, \psi\rangle$ also works since $\|\bar{\psi}_0^{\perp}\|_{L^2(dx)}$ is small. The orthogonal splitting $\psi = \bar{\psi}_0 \oplus \psi_0^{\perp}$ and the projection $\Pi(\cdot)$ in (8.25) imply

$$\begin{aligned} 0 &= \Pi\big(\mathcal{R}(\bar{\psi}_0 + \psi_0^{\perp})\big) \\ &= \Pi\big(\mathcal{R}(\bar{\psi}_0)\big) + \Pi\big(\mathcal{R}(\psi_0^{\perp})\big) \\ &= \Pi(\mathcal{R}\bar{\psi}_0) + \dot{\psi_0}^{\perp} + iM^{1/2}(\mathcal{V} - V_0)\psi_0^{\perp} + i\Pi\bigg(\frac{GM^{-1/2}}{2}\Delta_X(G^{-1}\psi_0^{\perp})\bigg), \end{aligned}$$

where the last step follows from the orthogonal splitting

$$\Pi\big((\mathcal{V} - V_0)\psi_0^{\perp}\big) = (\mathcal{V} - V_0)\psi_0^{\perp}$$

together with the second order change in the subspace projection

$$\psi_0^{\perp}(\tau + \Delta\tau) = \Pi(\tau + \Delta\tau)\big(\psi_0^{\perp}(\tau + \Delta\tau)\big) = \Pi(\tau)\big(\psi_0^{\perp}(\tau + \Delta\tau)\big) + \mathcal{O}(\Delta\tau^2)$$

which yields $\Pi(\dot{\psi_0^{\perp}}) = \dot{\psi_0^{\perp}}$; here $\Pi(\tau)\cdot$ denotes the projection on the orthogonal complement to the eigenvector $\bar{\psi}_0(\tau)$. To explain the second order change start with a function $v$ satisfying $\langle v, \Psi_{BO}(X_\tau)\rangle = 0$ and $\Psi_{BO}(X_\sigma) = \Psi_{BO}(X_\tau) + \mathcal{O}(\Delta\tau)$ for $\sigma \in [\tau, \tau + \Delta\tau]$ to obtain

$$\begin{aligned} \Pi(\sigma)\big(\Pi(\tau + \Delta\tau)v - \Pi(\tau)v\big) &= \Pi(\sigma)\Big(\langle v, \Psi_{BO}(X_\tau)\rangle \Psi_{BO}(X_\tau) - \langle v, \Psi_{BO}(X_{\tau+\Delta\tau})\rangle \Psi_{BO}(X_{\tau+\Delta\tau})\Big) \\ &= \Pi(\sigma)\mathcal{O}(\Delta\tau^2) + \Pi(\sigma)\Big(\langle v, \mathcal{O}(\Delta\tau)\rangle \Psi_{BO}(X_\tau)\Big) \\ &= \mathcal{O}(\Delta\tau^2) + \mathcal{O}(\Delta\tau)\Big(\Psi_{BO}(X_\tau) - \langle \Psi_{BO}(X_\tau), \Psi_{BO}(X_\sigma)\rangle \Psi_{BO}(X_\sigma)\Big) \\ &= \mathcal{O}(\Delta\tau^2). \end{aligned}$$

Let $\widetilde{\mathcal{S}}_{\tau,\sigma}$ be the solution operator from time $\sigma$ to $\tau$ for the generator

$$v \mapsto iM^{1/2}(\mathcal{V} - V_0)v + i\Pi\bigg(\frac{GM^{-1/2}}{2}\Delta_X(G^{-1}v)\bigg) =: iM^{1/2}\hat{\mathcal{V}}v\,.$$

Consequently, the perturbation $\psi_0^{\perp}$ can be determined from the projected residual

$$\dot{\psi}_0^{\perp} = -iM^{1/2}\hat{\mathcal{V}}\psi_0^{\perp} - \Pi(\mathcal{R}\bar{\psi}_0)$$

and we have the solution representation

$$\psi_0^{\perp}(\tau) = \widetilde{\mathcal{S}}_{\tau,0}\psi_0^{\perp}(0) - \int_0^{\tau} \widetilde{\mathcal{S}}_{\tau,\sigma}\Pi\big(\mathcal{R}\bar{\psi}_0(\sigma)\big)\,d\sigma\,. \tag{8.26}$$

Integration by parts introduces the factor $M^{-1/2}$ we seek

$$\begin{aligned} \int_0^{\tau} \widetilde{\mathcal{S}}_{\tau,\sigma}\Pi\mathcal{R}\bar{\psi}_0(\sigma)\,d\sigma &= \int_0^{\tau} iM^{-1/2}\frac{d}{d\sigma}(\widetilde{\mathcal{S}}_{\tau,\sigma})\hat{\mathcal{V}}^{-1}\Pi\mathcal{R}\bar{\psi}_0(\sigma)\,d\sigma \\ &= \int_0^{\tau} iM^{-1/2}\frac{d}{d\sigma}\Big(\widetilde{\mathcal{S}}_{\tau,\sigma}\hat{\mathcal{V}}^{-1}\Pi\mathcal{R}\bar{\psi}_0(\sigma)\Big)\,d\sigma \\ &\quad - \int_0^{\tau} iM^{-1/2}\widetilde{\mathcal{S}}_{\tau,\sigma}\frac{d}{d\sigma}\Big(\hat{\mathcal{V}}^{-1}(X_\sigma)\Pi\mathcal{R}\bar{\psi}_0(\sigma)\Big)\,d\sigma \\ &= iM^{-1/2}\hat{\mathcal{V}}^{-1}\Pi\mathcal{R}\bar{\psi}_0(\tau) - iM^{-1/2}\widetilde{\mathcal{S}}_{\tau,0}\hat{\mathcal{V}}^{-1}\Pi\mathcal{R}\bar{\psi}_0(0) \\ &\quad - \int_0^{t} iM^{-1/2}\widetilde{\mathcal{S}}_{\tau,\sigma}\frac{d}{d\sigma}\Big(\hat{\mathcal{V}}^{-1}(X_\sigma)\Pi\mathcal{R}\bar{\psi}_0(\sigma)\Big)\,d\sigma\,. \end{aligned} \tag{8.27}$$

To analyze the integral in the right hand side we will use the fact

$$\hat{\mathcal{V}}^{-1} = \Big(I + (\mathcal{V}-V_0)^{-1}\Big[\hat{\mathcal{V}} - (\mathcal{V}-V_0)\Big]\Big)^{-1}(\mathcal{V}-V_0)^{-1},$$

which can be verified by multiplying both sides from the left by $I + (\mathcal{V}-V_0)^{-1}\Big[\hat{\mathcal{V}} - (\mathcal{V}-V_0)\Big]$. A spectral decomposition in $L^2(dx)$, based on the electron eigenpairs $\{\lambda_k, \bar{\psi}_k\}_{k=1}^{\infty}$ and satisfying $\mathcal{V}\bar{\psi}_k = \lambda_k\bar{\psi}_k$, then implies

$$\begin{aligned}\hat{\mathcal{V}}^{-1}\Pi(\mathcal{R}\bar{\psi}_0) &= \Big(I + (\mathcal{V}-V_0)^{-1}\Big[\hat{\mathcal{V}} - (\mathcal{V}-V_0)\Big]\Big)^{-1}(\mathcal{V}-V_0)^{-1}\Pi(\mathcal{R}\bar{\psi}_0)\\ &= \sum_{k\neq 0}\Big(I + (\mathcal{V}-V_0)^{-1}\Big[\hat{\mathcal{V}} - (\mathcal{V}-V_0)\Big]\Big)^{-1}(\lambda_k - V_0)^{-1}\psi_k\langle\Pi(\mathcal{R}\bar{\psi}_0),\psi_k\rangle\\ &= \sum_{k\neq 0}(\lambda_k - V_0)^{-1}\psi_k\langle\Pi(\mathcal{R}\bar{\psi}_0),\psi_k\rangle + \mathcal{O}(M^{-1})\end{aligned} \tag{8.28}$$

which applied to the integral in the right hand side of (8.27) shows that $\|\bar{\psi}_0^{\perp}\|_{L^2(dx)} = \mathcal{O}(M^{-1/2})$ on a bounded time interval, when the spectral gap condition holds and $\psi_k$ are smooth.

The evolution on longer times requires an additional idea: one can integrate by parts recursively in (8.27) to obtain

$$\int_0^{\tau}\widetilde{\mathcal{S}}_{\tau,\sigma}\Pi\mathcal{R}\bar{\psi}_0(\sigma)\,d\sigma = \Big[\widetilde{\mathcal{S}}_{\tau,\sigma}\Big(\tilde{\mathcal{B}}\tilde{\mathcal{R}} - \tilde{\mathcal{B}}\frac{d}{d\sigma}(\tilde{\mathcal{B}}\tilde{\mathcal{R}}) + \tilde{\mathcal{B}}\frac{d}{d\sigma}\big(\tilde{\mathcal{B}}\frac{d}{d\sigma}(\tilde{\mathcal{B}}\tilde{\mathcal{R}})\big) - \dots\Big)\Big]_{\sigma=0}^{\sigma=\tau},$$
$$\tilde{\mathcal{B}} := iM^{-1/2}\hat{\mathcal{V}}^{-1}, \quad \tilde{\mathcal{R}} := \Pi\mathcal{R}\bar{\psi}_0(\sigma),$$

so that by (8.26) we have

$$\psi_0^{\perp}(\tau) = \widetilde{\mathcal{S}}_{\tau,0}\psi_0^{\perp}(0) - \Big[\widetilde{\mathcal{S}}_{\tau,\sigma}\Big(\tilde{\mathcal{B}}\tilde{\mathcal{R}} - \tilde{\mathcal{B}}\frac{d}{d\sigma}(\tilde{\mathcal{B}}\tilde{\mathcal{R}}) + \tilde{\mathcal{B}}\frac{d}{d\sigma}\big(\tilde{\mathcal{B}}\frac{d}{d\sigma}(\tilde{\mathcal{B}}\tilde{\mathcal{R}})\big) - \dots\Big)\Big]_{\sigma=0}^{\sigma=\tau}.$$

By choosing

$$\bar{\psi}_0^{\perp}(\sigma)\Big|_{\sigma=0} = -\Big(\tilde{\mathcal{B}}\tilde{\mathcal{R}}(\sigma) - \tilde{\mathcal{B}}\frac{d}{d\sigma}(\tilde{\mathcal{B}}\tilde{\mathcal{R}})(\sigma) + \tilde{\mathcal{B}}\frac{d}{d\sigma}\big(\tilde{\mathcal{B}}\frac{d}{d\sigma}(\tilde{\mathcal{B}}\tilde{\mathcal{R}})\big)(\sigma) - \dots\Big)\Big|_{\sigma=0}$$

we get

$$\bar{\psi}_0^{\perp}(\tau) = -\sum_{n=0}^{\infty}\tilde{\mathcal{B}}_0^n\mathcal{R}_0(\tau), \tag{8.29}$$

where $\tilde{\mathcal{B}}_0 := -iM^{-1/2}\hat{\mathcal{V}}^{-1}\frac{d}{d\tau}$ and $\mathcal{R}_0 := iM^{-1/2}\hat{\mathcal{V}}^{-1}\tilde{\mathcal{R}}$. We assume this expansion (8.29) is convergent in $L^2(dx)$ for each $\tau$, which follows from the smoothness estimate

$$\|\tilde{\mathcal{B}}_0^n\mathcal{R}_0(\tau)\|_{L^2(dx)} \to 0 \text{ as } n\to\infty \tag{8.30}$$

and (8.28).

The next step, verifying that also the non linear problem for $V_0$ works, is based on the contraction obtained from

$$V_0 - \lambda_0 = \frac{\langle\psi,(\mathcal{V}-\lambda_0)\psi\rangle}{\langle\psi,\psi\rangle} = \mathcal{O}(\|\psi_0^{\perp}\|_{L^2(dx)})$$

and that $\psi_0^{\perp}$ depends on $V_0$ in (8.26), (8.27) and (8.28) with a multiplicative factor $\mathcal{O}(M^{-1/2})$.

Finally, to conclude that $|\langle\psi,\psi\rangle - 1| = \mathcal{O}(M^{-1})$, we use the evolution equation

$$\frac{d}{dt}\langle\psi,\psi\rangle = M^{-1/2}|G|^2\mathrm{Im}\,\langle\Delta\frac{\psi}{G},\frac{\psi}{G}\rangle = \mathcal{O}(M^{-1})$$

where the last equality uses the obtained bound of $\psi_0^{\perp}$ in the first part of (8.23). The assumption of a finite hitting time $\tau^*$ then implies that $|\langle\psi,\psi\rangle - 1| = \mathcal{O}(\tau^*M^{-1}) = \mathcal{O}(M^{-1})$, since we may assume that $\langle\psi,\psi\rangle = 1$ on $I_\psi$. □

**Remark 8.3** (Error estimates for the densities)**.** We have the densities

$$\rho_{\mathrm{S}} = G_{\mathrm{S}}^{-2}\langle\psi,\psi\rangle \qquad \text{for the Schrödinger equation,} \tag{8.31}$$

$$\rho_{\mathrm{BO}} = G_{\mathrm{BO}}^{-2} \qquad \text{for the Born-Oppenheimer dynamics.} \tag{8.32}$$

From the stability of the Hamilton-Jacobi equation for $\log(|G|^{-2})$ and the estimate $\|\partial_{X^iX^j}(\theta-\tilde{\theta})\|_{L^\infty} = \mathcal{O}(M^{-1+\delta})$ in (8.18) we have

$$G_{\mathrm{S}}^{-2} = G_{\mathrm{BO}}^{-2} + \mathcal{O}(M^{-1+\delta})\,,$$

and Lemma 8.2 implies

$$\langle\psi,\psi\rangle = 1 + \mathcal{O}(M^{-1})\,, \tag{8.33}$$

which proves

$$\rho_{\mathrm{S}} = \rho_{\mathrm{BO}} + \mathcal{O}(M^{-1+\delta})\,.$$

## 9. Numerical examples

In order to demonstrate the presented theory we consider two different low dimensional Schrödinger problems. For both of these problems we show that there exists a Schrödinger eigenfunction density which converges weakly to the corresponding molecular dynamics density as $M \to \infty$ with a convergence rate within the upper bound predicted in the theoretical part of this paper.

9.1. **Example 1: A single WKB state.** The first problem we consider is the time-independent Schrödinger equation

$$\mathcal{H}\Phi := \left(-\frac{1}{2M}\partial_{XX} + \overline{\mathcal{V}}\right)\Phi = E\Phi \tag{9.1}$$

with heavy coordinate $X \in (-\pi,\pi]$ and two-state light coordinate $x \in \{x_-, x_+\}$. Periodicity is assumed over the heavy coordinate, $\Phi(X,x) = \Phi(X+2\pi,x)$, and the potential operator $\overline{\mathcal{V}}$ is defined by the matrix

$$\overline{\mathcal{V}}(X) = \begin{bmatrix} V(X) & \frac{1}{2}V(X)e(X)+c \\ \frac{1}{2}V(X)e(X)+c & 0 \end{bmatrix}, \tag{9.2}$$

where we have chosen $V(X) = -2\cos(X) + \cos(4X)$, $e(X) = 1 + X^2$ and $c$ to be a non-negative constant relating to the size of the spectral gap of $\overline{\mathcal{V}}$. The action $\overline{\mathcal{V}}\Phi$ is thus defined by

$$(\overline{\mathcal{V}}\Phi)(X,\cdot) \equiv \overline{\mathcal{V}}(X)\begin{pmatrix} \Phi(X,x_-) \\ \Phi(X,x_+) \end{pmatrix}.$$

For each $X$ the potential matrix (9.2) gives rise to the eigenvalue problem

$$\overline{\mathcal{V}}(X)\upsilon = \lambda_\pm(X)\upsilon$$

with the eigenvalues

$$\lambda_\pm(X) = \frac{1}{2}\left(V(X) \pm \mathrm{Sgn}(X)\sqrt{V(X)^2 + 4\big(V(X)e(X)/2+c\big)^2}\right),$$

where $\mathrm{Sgn}(X) = \pm 1$ as defined below. When constructing the molecular dynamics density for this problem

$$\rho_{\mathrm{MD}}(X) = \frac{C}{\sqrt{2(E-\lambda(X))}}\,,$$

one has to determine on which of the two eigenfunctions $\lambda_\pm$ to base this density. When $c = 0$ the difficulty that the eigenvalue functions $\lambda_+$ and $\lambda_-$ can cross is added to the problem. In order to determine the continuation of eigenvalue functions at the crossings we introduce a function $\mathrm{Sgn}(X)$ which is a sign function with $\mathrm{Sgn}(-\pi) = 1$ that changes sign at points where

$$V(X)^2 + 4\left(\frac{1}{2}V(X)e(X)+c\right)^2 = 0\,.$$

Since this situation can only occur when $c = 0$, it is possible to set

$$\mathrm{Sgn}(X) := \mathrm{sgn}(V(-\pi))\mathrm{sgn}(V(X))\,.$$

See Figure 8 for a typical eigenvalue function crossing, which makes the function $\lambda_\pm : \mathbb{R} \to \mathbb{R}$ smooth (in contrast to the choice Sgn $\equiv 1$).

To solve (9.1) numerically, we use the finite difference method to discretise the operator $\mathcal{H}$ on a grid $\{X_j\}_{j=1}^N \times \{x_-, x_+\}$ with the step-size $h = 2\pi/N$ and $X_j = jh$. The discrete eigenvalue problem

$$\mathcal{H}^{(h)}\Upsilon_j = E_j\Upsilon_j$$

is solved for the 10 eigenvalues being closest to the fixed energy $E$ and a molecular dynamics approximation of the eigensolution is constructed by

$$\Phi_{\mathrm{MD}}(X, x) := \sqrt{\rho_{\mathrm{MD}}(X)}\, e^{iM^{1/2}\Theta(X)} v(X, x)\,,$$

where $v(X, \cdot)$ is one of the eigenvectors of $\overline{\mathcal{V}}(X)$ and

$$\Theta(X) := \int_0^X \sqrt{2(E_1 - \lambda(s))}\, ds \tag{9.3}$$

is approximated by a trapezoidal quadrature yielding $\Theta^{(h)}$. Thereafter a Schrödinger eigensolution $\Phi^{(h)}$ which is close to the molecular dynamics eigensolution is obtained by projecting $\Phi_{MD}$ onto the subspace spanned by $\{\Upsilon\}_{j=1}^{\bar{J}}$ as described in Algorithm 2. By denoting $\rho_{\Phi^{(h)}}(X) = \langle \Phi^{(h)}, \Phi^{(h)} \rangle$ and $\rho_{\mathrm{MD}}(X) = \langle \Phi_{\mathrm{MD}}, \Phi_{\mathrm{MD}} \rangle$, the observables $g_1(X) = X^2$ and $g_2(X) = V(X)$ are used to compute the convergence rate of

$$\left| \frac{\int_{-\pi}^{\pi} g_i(X)\rho_{\mathrm{MD}}(X)\, dX - \int_{-\pi}^{\pi} g_i(X)\rho_{\Phi^{(h)}}(X)\, dX}{\int_{-\pi}^{\pi} g_i(X)\rho_{\mathrm{MD}}(X)\, dX} \right|, \tag{9.4}$$

as $M$ increases. Further details of the numerical solution idea are described in Algorithm 1.

Plots of the results for the test case with the spectral gap $c = 5$ and $E = 0$, and for the test case with crossing eigenvalue functions when $c = 0$ and $E = 1.2$ are given below. Most noteworthy is Figure 11, which demonstrates that the obtained convergence rate for (9.4) is $\mathcal{O}(M^{-1})$ for both scenarios.

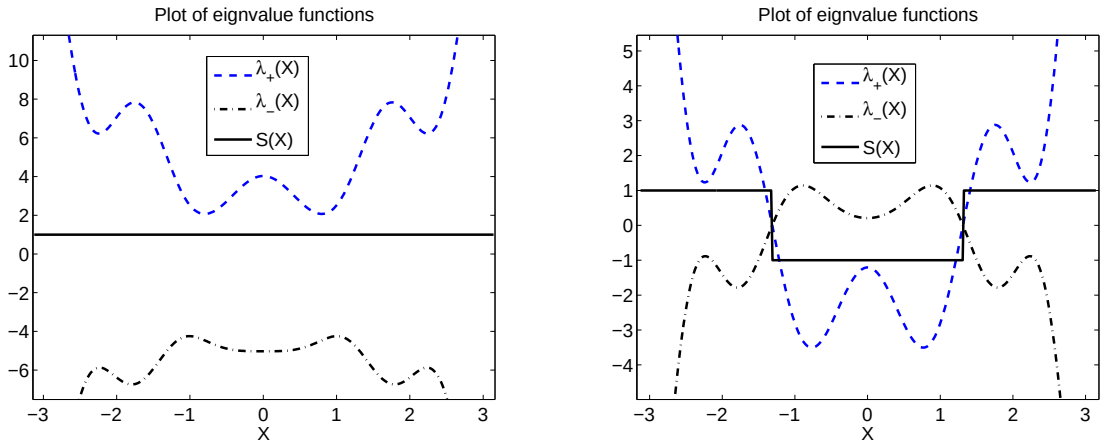


FIGURE 8. Left plot: Eigenvalue functions when $c = 5$. There is a spectral gap which makes the sign function constant $S = 1$. Right plot: Eigenvalue functions when $c = 0$. The eigenvalue functions exhibit crossing, consequently the function $S$ changes its sign from $\pm 1$ to $\mp 1$ at the crossing points.

9.2. **Example 2: A caustic state.** Next, we consider the one dimensional, time independent, periodic Schrödinger equation

$$\left(-\frac{1}{2M}\partial_{XX} + V\right)\Phi = E\Phi\,, \quad X \in (-2\sqrt{E}, 2\sqrt{E}) \tag{9.7}$$

**Algorithm 1** Algorithm for problems in Example 1

**Input:** Energy $E$; potential functions $V$, $e$ and $c$; mass $M$; number of grid points $N$ and grid $\{X_i\}_{i=1}^N$.
**Output:** Schrödinger projection density $\rho_{\Phi^{(h)}}$.

**1.** Construct the discrete operator $\mathcal{H}^{(h)}$ from (9.1) using finite differences and solve the eigenvalue problem

$$\mathcal{H}^{(h)}\Upsilon_i = E_i\Upsilon_i$$

for the 10 eigenvalues being closest to $E$ by using MATLAB **eigs(H,10,E)**.

**2.** Sort the eigenvalues and eigenvectors by distance from $E$ and keep only the $E_i$s which are less than $M^{-1/2}$ away from $E$. Let $\bar{J}$ be the number of kept eigenvalues and $E_0$ the eigenvalue closest to $E$.

**3.**
**for** $i = 1$ to $N$ **do**
Solve the eigenvalue problem

$$\overline{\mathcal{V}}(X_i,\cdot)\upsilon_\pm(X_i,\cdot) = \lambda_\pm(X_i)\upsilon_\pm(X_i,\cdot)\,,$$

where $\overline{\mathcal{V}}$ is the matrix defined in (9.2).
**end for**

**4.** Construct the molecular dynamics density according to the formula

$$\rho_{\mathrm{MD}}(X) = \frac{(E_0-\lambda(X))^{-1/2}}{\int_{[0,2\pi]}(E_0-\lambda(X))^{-1/2}\,dX}\,,$$

where we choose $\lambda(X)$ above from the two eigenvalues $\lambda_\pm(X)$ by the criterion that the eigenvalue chosen must fulfil $\|\lambda\|_\infty < E_0$.

**5.** Construct a discrete molecular dynamics approximation to the eigenfunction

$$\Phi_{\mathrm{MD}}(X,x) = \sqrt{\rho_{\mathrm{MD}}(X)}e^{iM^{1/2}\Theta(X)}\upsilon(X,x)\,, \tag{9.5}$$

where $\upsilon(X,x)$ is one of the eigenvectors $\upsilon_\pm$,

$$\Theta(X) := \int_0^X \sqrt{2(E_1-\lambda(s))}\,ds\,, \tag{9.6}$$

and we approximate $\Theta$ by a trapezoidal quadrature $\Theta^{(h)}$.

**6.** Project the molecular dynamics solution $\Phi_{\mathrm{MD}}$ onto the eigenspace $\{\Upsilon_i\}_{i=1}^{\bar{J}}$, $\bar{J}\le 10$ by Algorithm 2 to obtain a projection solution $\Phi^{(h)}$.

**7.** Derive the Schrödinger projection density by
**for** $i = 1$ to $N$ **do**

$$\rho_{\Phi^{(h)}}(X_i) = |\Phi^{(h)}(X_i,x_-)|^2 + |\Phi^{(h)}(X_i,x_+)|^2\,,$$

**end for**
and scaling $\rho_{\Phi^{(h)}} = \rho_{\Phi^{(h)}}/\|\rho_{\Phi^{(h)}}\|$.

with $V(X) = X^2$ and $E = 1$. The eikonal equation corresponding to (9.7) is

$$\frac{1}{2}P^2 + V(X) = E\,. \tag{9.8}$$

As in Example 1, we would like to use the eikonal equation to construct a numerical approximate solution of (9.7) whose density converges weakly as $M\to\infty$ to the density generated from a solution of (9.7). The molecular dynamics density corresponding to this eikonal equation becomes by (3.19) $\rho_{\mathrm{BO}} = C(E - V(X))^{-1/2}$. The density $\rho_{\mathrm{BO}}$ goes to infinity at the caustics $X = V^{-1}(E) = \pm\sqrt{E}$ and the approach in Example 1 does not work directly. We will instead construct the numerical approximate solution using the stationary phase method as outlined below based on the WKB Fourier integral ansatz.

**Algorithm 2** Projection algorithm

**Input:** Mass $M$; wave solution $\Phi$; eigenvalues $\{E_i\}_{i=1}^{\bar{J}}$ and corresponding eigenvectors $\{\Upsilon_i\}_{i=1}^{\bar{J}}$.
**Output:** Schrödinger projection wave solution $\Phi^{(h)}$.

**1.** Organize eigenvalues by multiplicity by a numerical approximation. Construct a $\bar{J} \times \bar{J}$, zero matrix $A$ which keeps track of multiplicity relations as follows:
**for** $i = 1$ to $\bar{J}$ **do**
  **for** $j = i$ to $\bar{J}$ **do**
    **if** $|E_i - E_j| < M^{-3/4}$ **then**
      Consider eigenvalues equal since the expected spectral gap is $\mathcal{O}(M^{-1/2})$, and store this relation by
      **if** $A_{kj} = 0$ for all $k < i$ **then**
        Set $A_{ij} = 1$.
      **end if**
    **end if**
  **end for**
**end for**

**2.** For vectors $b \in \{0,1\}^{\bar{J}}$, define the projection

$$\Phi^{(h,b)} := \sum_{j,k=1}^{\bar{J}} b_k A_{k,j} \langle\langle \Phi, \Upsilon_j \rangle\rangle \Upsilon_j$$

and, letting $\rho$ and $\rho_{\Phi^{(h,b)}}$ denote the densities generated by $\Phi$ and $\Phi^{(h,b)}$ respectively, set

$$b^* = \arg \min_{b \in \{0,1\}^{\bar{J}}} \|\rho - \rho_{\Phi^{(h,b)}}\|.$$

**3.** Return the projection $\Phi^{(h)} := \Phi^{(h,b^*)}$.

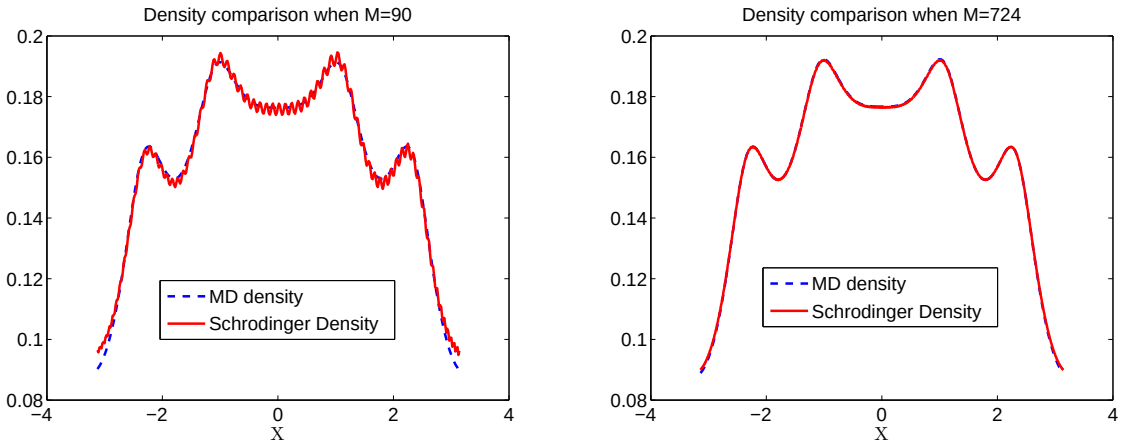


FIGURE 9. Plot of the MD density $\rho_{\text{MD}}$ and the Schrödinger projection density $\rho_{\Phi^{(h)}}$ in the case $c = 5$ and $E = 0$ for the two different masses $M = 90$ (left plot) and $M = 724$ (right plot) illustrating the convergence of the densities.

By the Legendre transform

$$\theta^*(P) = \min_X \big(XP - \theta(X)\big)$$

an invertible mapping between the momentum and position coordinates fulfilling $X = \nabla_P \theta^*(P)$ is constructed. Using equation (9.8), one sees that $\nabla_P \theta^*(P) = V^{-1}(E - P^2/2)$. Since $\theta^*(0) = 0$, one can derive

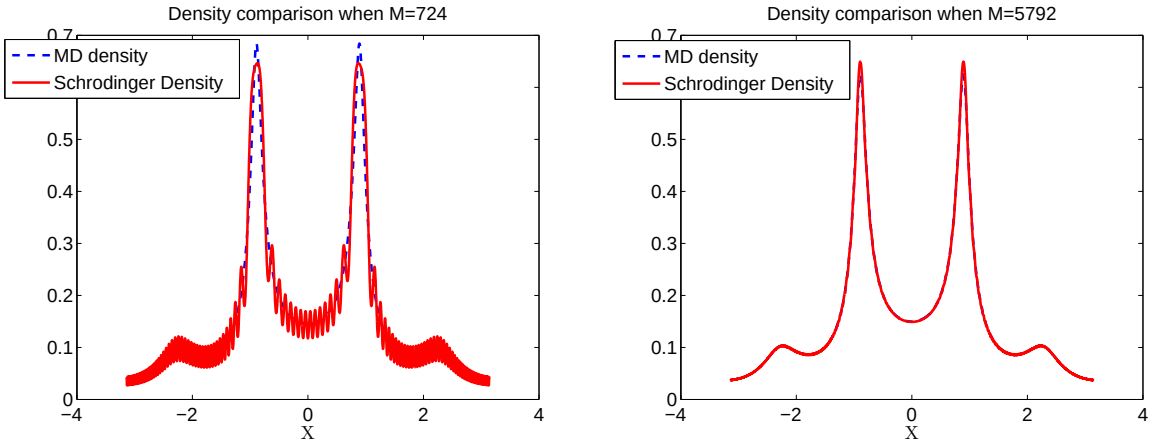


FIGURE 10. Plot of the MD density $\rho_{\mathrm{MD}}$ and Schrödinger projection density $\rho_{\Phi^{(h)}}$ in the case $c = 0$ and $E = 1.2$ for the two different masses $M = 724$ (left plot) and $M = 5792$ (right plot).

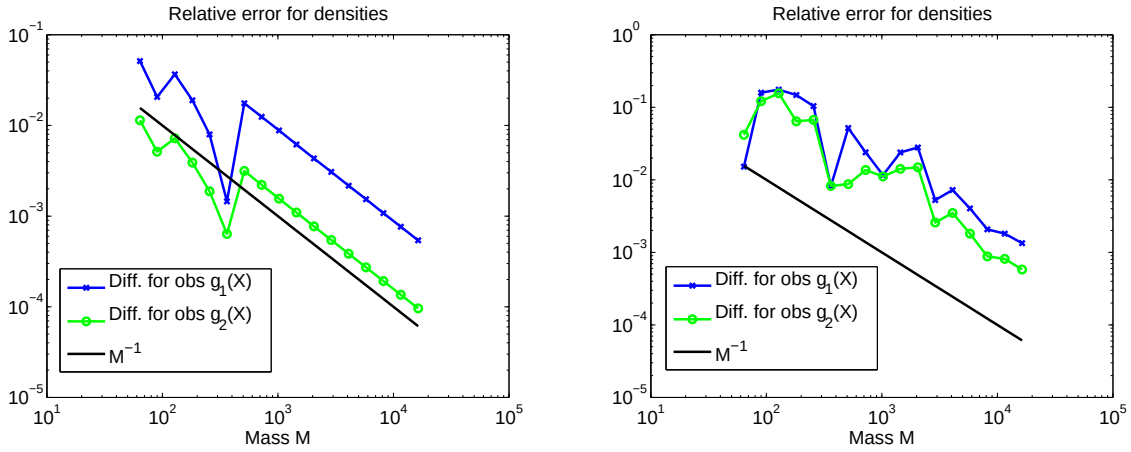


FIGURE 11. Left plot: Plot of the observable density errors given in (9.4) with an eigenvalue gap, when $c = 5$ and $E = 0$. Right plot: Plot of the observable density errors given in (9.4) with an eigenvalue crossing, when $c = 0$ and $E = 1.2$.

that for this particular choice of $V$

$$\theta^*(P) = \int_0^P \sqrt{E - s^2/2}\, ds = \frac{E}{\sqrt{2}} \left[ \sin^{-1}\left( \frac{P}{\sqrt{2E}} \right) + \frac{P}{\sqrt{2E}} \sqrt{1 - \frac{P^2}{2E}} \right].$$

In neighbourhoods of the caustics $[-2E^{1/2}, -X_0)$ and $(X_0, 2E^{1/2}]$, we construct the approximate solution by

$$\Phi(X) = \frac{u(X)}{\sqrt{|\nabla_X V(X)|}}$$

where $u$ is the inverse Fourier transform

$$u(X) := \int_{-2\sqrt{E}}^{2\sqrt{E}} e^{iM^{1/2}(-XP + \theta^*(P))}\, dP$$

and $X_0 \in (-V^{-1}(E), V^{-1}(E))$ is a value yet to be chosen. In the region $(-X_0, X_0)$ the approximate solution is constructed by

$$\Phi(X) = C\frac{\overline{u}(X)}{(E-V(X))^{1/4}}\,. \tag{9.9}$$

Here

$$\overline{u}(X) := e^{-iM^{1/2}\theta(X)}\psi_+ + e^{iM^{1/2}\theta(X)}\psi_-\,, \tag{9.10}$$

with, according to the Legendre transform, $\theta(X) := X\sqrt{2(E-V(X))} - \theta^*\left(\sqrt{2(E-V(X))}\right)$ and $\psi_\pm$ determined by the stationary phase method:

**1.** Set $P(p) = P_0 + p$ with $P_0 = \sqrt{2(E-V(X_0))}$ and let

$$Y(p) := \mathrm{sgn}(p)\sqrt{2\frac{-X(P_0+p)+\theta^*(P_0+p)+\theta(X_0)}{\partial_{PP}\theta^*(P_0)}}\,,$$

using

$$\theta(X) := X\sqrt{2(E_0-V(X))} - \theta^*\left(\sqrt{2(E_0-V(X))}\right)\,, \tag{9.11}$$

and determine its inverse $p(Y)$ in a neighbourhood of $Y=0$ by computing $(p_i, Y(p_i))$ on a grid around $p=0$ and, for $k \geq 3$, fit a $3k+1$th degree polynomial to the values $(Y(p_i), p_i)$ using the method of least squares.

**2.** Evaluate the stationary phase expansion

$$\begin{aligned} u(X_0) = \sum_{p_0=\pm\sqrt{2(E-V(X_0))}} & e^{i\pi\mathrm{sgn}(\partial_{PP}\theta^*(P_0))/4}\left[\left|\frac{1}{2}\partial_{PP}\theta^*(P_0)\right|^{-1/2} e^{-iM^{1/2}\theta(X_0)}\right. \\ & \left.\times\sum_{j=0}^{k}\frac{M^{-j/2}}{j!}\left(i\left(\frac{1}{2}\partial_{PP}\theta^*(P_0)\right)^{-1}\partial_{YY}\right)^j |\partial_Y p|\Big|_{Y=0} + \mathcal{O}(M^{-j/2})\right] \end{aligned} \tag{9.12}$$

to obtain

$$u(X_0^-) = e^{iM^{1/2}\theta(X_0)}(\psi_+ + \mathcal{O}(M^{-k/2})) + e^{-iM^{1/2}\theta(X_0)}(\psi_- + \mathcal{O}(M^{-k/2}))\,,$$

where

$$\psi_\pm := e^{i\pi\mathrm{sgn}(\partial_{PP}\theta^*(\pm P_0))/4}\left|\frac{1}{2}\partial_{PP}\theta^*(\pm P_0)\right|^{-1/2}\sum_{k=0}^{3}\frac{M^{-k/2}}{k!}\left(i\left(\frac{\partial_{PP}\theta^*(\pm p_0)}{2}\right)^{-1}\partial_{YY}\right)^k |\partial_Y p|\Big|_{Y=0}.$$

The constant $C$ in (9.9) is chosen so that the wave solution parts are continuous at the gluing point, $\Phi(\pm X_0^-) = \Phi(\pm X_0^+)$. It is most easy to determine $C$ when $X_0$ is chosen so that $|u(X_0)|$ is at a local maximum; see Figure 12 for an illustration of the gluing procedure.

At the end a Schrödinger eigenfunction solution $\Phi^{(h)}$ is obtained by projecting $\Phi$ onto the space spanned by a set of eigensolutions to the discretized version of the Schrödinger problem, $\{\Upsilon_j\}_{j=1}^{\bar{J}}$, as is described in Algorithm 2.

Two convergence results are needed to make the method work. First, the density generated from the stationary phase based on the approxmiate solution $\rho(X) := |\Phi|^2(X)/\|\Phi\|_2^2$ must converge weakly to the Schrödinger projection based density $\rho_{\Phi^{(h)}}(X) := |\Phi^{(h)}|^2(X)/\|\Phi^{(h)}\|_2^2$ as $M \to \infty$; see Figure 13 for an illustration of how these functions converge. Second, $\rho_{\Phi^{(h)}}$ must converge to the molecular dynamics density $\rho_{\mathrm{MD}}(X) := C(E-V(X))^{-1/2}$ as $M$ increases; see Figure 14.

A numerical test of the convergence rate of

$$\left|\frac{\int_{-2\sqrt{E_0}}^{2\sqrt{E_0}} g_1(X)\rho_{\mathrm{MD}}(X)\,dX}{\int_{-2\sqrt{E_0}}^{2\sqrt{E_0}} g_2(X)\rho_{\mathrm{MD}}(X)\,dX} - \frac{\int_{-2\sqrt{E_0}}^{2\sqrt{E_0}} g_1(X)\rho_{\Phi^{(h)}}(X)\,dX}{\int_{-2\sqrt{E_0}}^{2\sqrt{E_0}} g_2(X)\rho_{\Phi^{(h)}}(X)\,dX}\right| \tag{9.13}$$

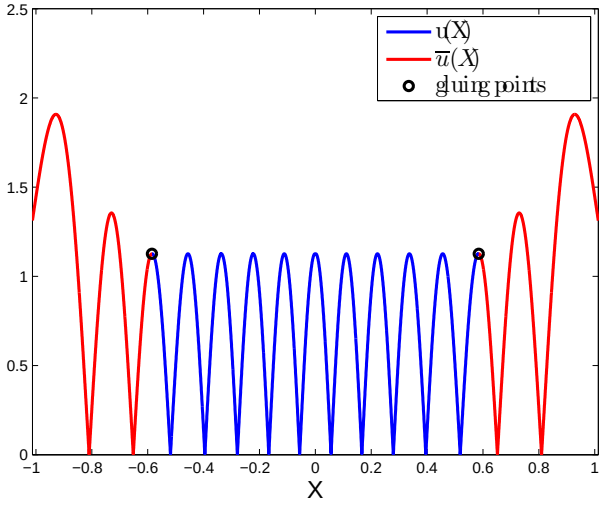


FIGURE 12. Plot illustrating the gluing procedure of the functions $u(X)$ and $\overline{u}(X)$ at the points $\pm X_0$.

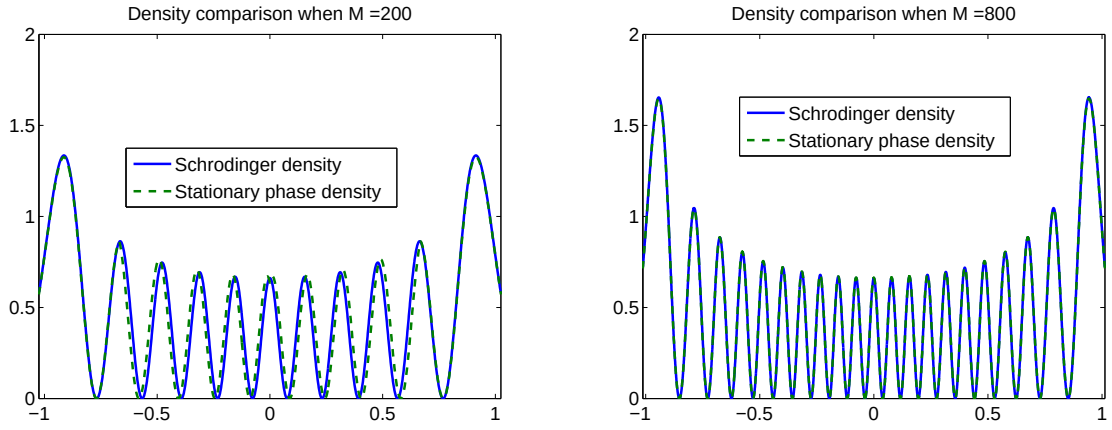


FIGURE 13. Comparison of the approximate solution based density $\rho$ and the Schrödinger projection based solution $\rho_{\Phi^{(h)}}$ for $M = 200$ (left plot) and $M = 800$ (right plot).

as $M$ increases is illustrated in Figure 15 for the observables

$$g_1(X) = \frac{(1.5-X)^6(1.5+X)^6(1+e^{-X^2})}{1.5^{12}} \quad \text{and} \quad g_2(X) = \frac{(1.5-X)^6(1.5+X)^6(1-X^2+X^4)}{1.5^{12}}\,. \tag{9.14}$$

Further details of the solution procedure in Exampe 2 are given in Algorithm 3.

## 10. The stationary phase expansion

Consider the phase function $\check{X}\cdot\check{P}-\theta^*(\hat{X},\check{P})$ and let $\check{P}_0(\hat{X})$ be any solution to the stationary phase equation $\check{X}=\nabla_{\check{P}}\theta^*(\hat{X},\check{P}_0)$. We rewrite the phase function

$$\check{X}\cdot\check{P}-\theta^*(\check{X},\check{P}) = \underbrace{\check{X}\cdot\check{P}_0-\theta^*(\check{X},\check{P}_0)}_{=\theta(\hat{X},\check{X})} + (\check{P}-\check{P}_0)\cdot\int_0^1 (1-t)\partial_{PP}\theta^*\left(\check{P}_0+t[\check{P}-\check{P}_0]\right)\,dt\,[\check{P}-\check{P}_0]\,.$$

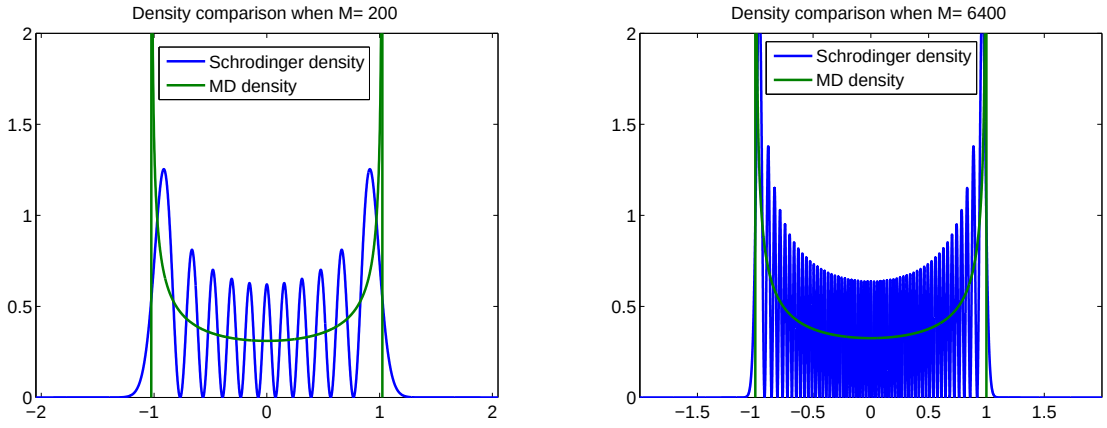


FIGURE 14. Comparision of the Schrödinger projection density $\rho_{\Phi^{(h)}}$ and the molecular dynamics density $\rho_{\mathrm{MD}}$ for $M = 200$ (left plot) and $M = 6400$ (right plot).

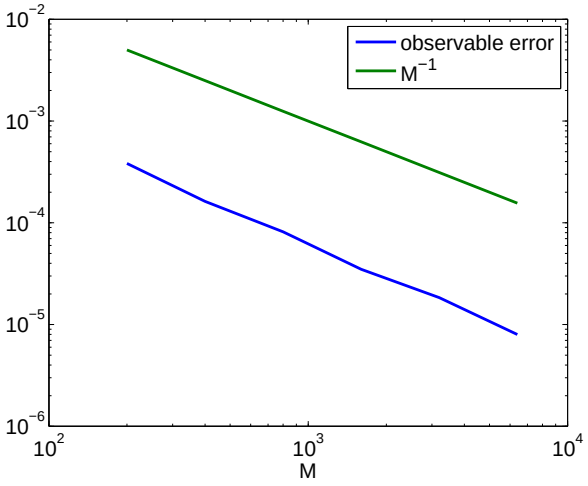


FIGURE 15. Convergence rate of (9.13) for the observables $g_1$ and $g_2$ as defined in (9.14).

The relation

$$\frac{1}{2} Y \cdot \partial_{PP}\bar{\theta}(\check{P}_0) Y = (\check{P} - \check{P}_0) \cdot \int_0^1 (1-t)\partial_{PP}\bar{\theta}\left(\check{P}_0 + t[\check{P} - \check{P}_0]\right)\, dt\, [\check{P} - \check{P}_0]$$

defines the function $Y(\check{P})$, and its inverse $\check{P}(Y)$, so that the phase is a quadratic function in $Y$. The stationary phase expansion of an integral takes the form, see [10],

$$\begin{aligned} &\int_{\mathbb{R}^d} w(\check{P})\, e^{-iM^{1/2}(\check{X}\cdot\check{P} - \theta^*(\hat{X},\check{P}))}\, d\check{P} \\ &\simeq \sum_{\nabla_P\theta^*(\check{P}_0)=\check{X}} (2\pi M^{-1/2})^{d/2} \left|\det \frac{\partial(\check{P})}{\partial(\check{X})}\right|^{1/2} e^{i\frac{\pi}{4}\mathrm{sgn}(\partial_{PP}\theta^*(\check{P}_0))}\, e^{-iM^{1/2}\theta(\hat{X},\check{X})} \\ &\quad \times \sum_{k=0}^{\infty} \frac{M^{-k/2}}{k!} \left(\sum_{l,j} i(\partial_{P^lP^j}\theta^*)^{-1}(\check{P}_0)\partial_{Y^lY^j}\right)^k \left(w(\check{P}(Y)) \left|\det \frac{\partial(\check{P})}{\partial(Y)}\right|\right)\Big|_{Y=0}. \end{aligned} \tag{10.1}$$

**Algorithm 3** Algorithm for Example 2

**Input:** An energy $E$, an one-dimensional potential function $V$, mass $M$, Schrödinger equation (9.7).
**Output:** The Schrödinger projection density $\rho_{\Phi^{(h)}}$.

**1.** Identify the right caustic point $X_+ > 0$ satisfying $X_+ = V^{-1}(E)$. For a fixed $E \in \mathbb{R}$, consider the periodic eigenvalue problem. Solve (9.7) numerically by constructing the discretised operator form of $-(2M)^{-1}\partial_{XX} + V$ using finite differences and denoted $\mathcal{H}^{(h)}$, and solve the eigenvalue problem

$$\mathcal{H}^{(h)} P_i = E_i P_i \tag{9.15}$$

for the 10 eigenvalues closest to $E$ using the Matlab eigenvalue solver **eigs(H,10,E)**. Let $E_0$ denote the eigenvalue closest to $E$ and consider from now on solving (9.7) for the energy $E_0$ and its corresponding eikonal equation $\frac{1}{2}P^2 + V(X) = E_0$.

**2.** Determine $\theta^*(P)$ by

$$\theta^*(P) = \int_0^P \nabla_P \theta^*(p)\, dp$$

**3.** Evaluate the Fourier integral

$$u(X) := \int_{-2\sqrt{E}}^{2\sqrt{E}} e^{iM^{1/2}(-XP+\theta^*(P))}\, dP\,, \quad |X| > X_0\,, \tag{9.16}$$

where $X_0$ is chosen as the smallest value $X > X_+/2$ such that $|u(X)|$ is at a local maximum, and for $|X| \le X_0$ compute $\overline{u}$ by (9.10) using the stationary phase method.

**4.** Construct the approximate solution

$$\Phi(X) := \begin{cases} C\overline{u}(X)(E_0 - V(X))^{-1/4} & |X| \le X_0\,, \\ u(X)/\sqrt{|\nabla_X V(X)|} & |X| \ge X_0\,, \end{cases}$$

with

$$C = \frac{u(X_0)(E_0 - V(X_0))^{1/4}}{\sqrt{|\nabla_X V(X_0)|}\overline{u}(X_0)}\,.$$

**5.**
Project $\Phi$ onto the eigenspace $\{\Upsilon_i\}_{i=1}^{\bar{J}}$, $\bar{J} \le 10$ by Algorithm 2 to obtain a projection solution $\Phi^{(h)}$ and compute its corresponding approximate density

$$\rho_{\Phi^{(h)}} = \frac{|\Phi^{(h)}|^2(X)}{\|\Phi^{(h)}\|_2^2}\,.$$


## Acknowledgment

The research of P.P. and A.S. was partially supported by the National Science Foundation under the grant NSF-DMS-0813893 and Swedish Research Council grant 621-2010-5647, respectively. P.P. also thanks KTH and Nordita for their hospitality during his visit when the presented research was initiated.


## References


[1] F.A. Berezin and M.A. Shubin, *The Schrödinger equation*, Kluwer Academic Publishers, 1991.
[2] M. Born and R. Oppenheimer, *Zur quantentheorie der molekeln*, Ann. Physik (1927), no. 84, 4571–484.
[3] F.A. Bornemann, P. Nettesheim, and C. Schütte, *Quantum-classical molecular dynamics as an approximation to full quantum dynamics*, J. Chem. Phys. **105** (1996), 1074–1083.
[4] A. Bouzounia and D. Robert, *Uniform semiclassical estimates for the propagation of quantum observables*, Duke Math. J. **111** (2002), 223–252.
[5] J. Briggs, S. Boonchui, and S. Khemmani, *The derivation of the time-dependent Schrödinger equation*, J. Phys. A: Math. Theor. **40** (2007), 1289–1302.
[6] J. Briggs and J.M. Rost, *On the derivation of the time-dependent equation of Schrödinger*, Foundations of Physics **31** (2001), 693–712.

[7] E. Cances, M. Defranceschi, W. Kutzelnigg, C. LeBris, and Y. Maday, *Computational chemistry: a primer*, Handbook of Numerical Analysis, vol. X, North-Holland, 2007.
[8] E. Cances, F. Legoll, and G. Stolz, *Theoretical and numerical comparison of some sampling methods for molecular dynamics*, Math. Model. Num. Anal. **41** (2007), 351–389.
[9] J. Carlsson, M. Sandberg, and A. Szepessy, *Symplectic Pontryagin approximations for optimal design*, Math. Model. Num. Anal. **43** (2009), 3–32.
[10] J.J. Duistermaat, *Fourier integral operators*, Courant Institute, 1973.
[11] J.-P. Eckmann and R. Sénéor, *The Maslov-WKB Method for the (an-)harmonic oscillator*, Arch. Rat. Mech. Anal. **61** (1976), 153–173.
[12] L.C. Evans, *Partial differential equation*, American Mathematical Society, Providence, RI, 1998.
[13] C. Fefferman and L. Seco, *Eigenvalues and eigenfunctions of ordinary differential operators*, Adv. Math. **95** (1992), 145–305.
[14] D. Frenkel and B. Smith, *Understanding molecular simulation*, Academic Press, 2002.
[15] G.A. Hagedorn, *High order corrections to the time-independent Born-Oppenheimer approximation II: diatomic Coulomb systems*, Comm. Math. Phys. **116** (1988), 23–44.
[16] B. Helffer, *Semi-classical analysis for the Schrödinger operator and applications*, Lecture Notes in Mathematics, vol. 1336, Springer Verlag, 1988.
[17] H. Jeffreys, *On certain approximate solutions of linear differential equations of the second order*, Proc. London Math. Soc. **23** (1924), 428–436.
[18] J. B. Keller, *Corrected Bohr-Sommerfeld quantum conditions for nonseparable systems*, Ann. Phys. **4** (1958), 180–188.
[19] M. Klein, A. Martinez, R. Seiler, and X. P. Wang, *On the Born-Oppenheimer expansion for polyatomic molecules*, Comm. Math. Phys. **143** (1992), 607–639.
[20] C. Lasser and S. Röblitz, *Computing expectations values for molecular quantum dynamics*, SIAM J. Sci. Comput. **32** (2010), 1465–1483.
[21] C. LeBris, *Computational chemistry from the perspective of numerical analysis*, Acta Numerica, vol. 14, pp. 363–444, CUP, 2005.
[22] E. Lieb and R. Seiringer, *The stability of matter in quantum mechanics*, CUP, 2010.
[23] D. Marx and J. Hutter, *Ab initio molecular dynamics: Theory and implementation, modern methods and algorithms of quantum chemistry*, Tech. report, John von Neumann Institute for Computing, Jülich, 2001.
[24] A. Martinez and V. Sordoni, *Twisted pseudodifferential calculus and application to the quantum evolution of molecules*, Memoirs Am. Math. Soc., **200** (2009), n. 936.
[25] V. P. Maslov and M. V. Fedoriuk, *Semi-classical approximation in quantum mechanics*, D. Reidel Publishing Company, 1981; based on: V. P. Maslov, *Theory of perturbations and asymptotic methods*, Moskov. Gos. Univ.. Moscow 1965 (Russian).
[26] M.Dimassi and J. Sjöstrand, *Spectral asymptotics in the semiclassical limit*, LMS Lecture Note Series, vol. 268, CUP, 1999.
[27] N. F. Mott, *On the theory of excitation by collision with heavy particles*, Proc. Camb. Phil. Soc. **27** (1931), 553–560.
[28] G. Panati, H. Spohn, and S. Teufel, *Space-adiabatic perturbation theory*, Adv. Theor. Math. Phys. **7** (2003), 145–204.
[29] Rayleigh, *On the propagation of waves through a stratified medium, with special reference to the question of reflection*, Proc. Roy. Soc. (London) Series A **86** (1912), 207–226.
[30] M. Reed and B. Simon, *Methods of Modern Mathematical Physics, I: Functional Analysis*, Academic Press INC., 1980.
[31] M. Sandberg and A. Szepessy, *Convergence rates of symplectic Pontryagin approximations in optimal control theory*, Math. Model. Num. Anal. **40** (2006), 149–173.
[32] L. Schiff, *Quantum mechanics*, McGraw-Hill, 1968.
[33] E. Schrödinger, *Collected papers on wave mechanics*, Blackie and Son, London, 1928.
[34] A. Szepessy, *Langevin molecular dynamics derived from Ehrenfest dynamics*, Math. Mod. Meth. Appl. S. **21** (2011), 2289-2334.
[35] D. J. Tanner, *Introduction to quantum mechanics: A time-dependent perspective*, University Science Books, 2006.
[36] J. C. Tully, *Mixed quantum-classical dynamics*, Faraday Discuss. **110** (1998), 407–419.
[37] E. von Schwerin and A. Szepessy, *A stochastic phase-field model determined from molecular dynamics*, Math. Model. Num. Anal. **44** (2010), 627–646.
[38] S. Zelditch, *Quantum ergodicity and mixing of eigenfunctions*, in Franoise, Jean-Pierre; Naber, Gregory L.; Tsun, Tsou Sheung, Encyclopedia of mathematical physics. Vol. 1, 2, 3, 4, 5, Academic Press/Elsevier Science, Oxford, 2006.

Department of Mathematics University of Vienna Nordbergstrae 15 1090 Wien, Austria
*E-mail address*: christian.bayer@univie.ac.at

Department of Numerical Analysis, Kungl. Tekniska Högskolan, 100 44 Stockholm, Sweden
*E-mail address*: hhoel@kth.se

Department of Numerical Analysis, Kungl. Tekniska Högskolan, 100 44 Stockholm, Sweden
*E-mail address*: smakadir@csc.kth.se

Department of Mathematical Sciences, University of Delaware, Newark, DE 19716, USA
*E-mail address*: plechac@math.udel.edu

Department of Numerical Analysis, Kungl. Tekniska Högskolan, 100 44 Stockholm, Sweden
*E-mail address*: msandb@kth.se

Department of Mathematics, Kungl. Tekniska Högskolan, 100 44 Stockholm, Sweden
*E-mail address*: szepessy@kth.se

Division of Mathematics, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia
*E-mail address*: raul.tempone@kaust.edu.sa