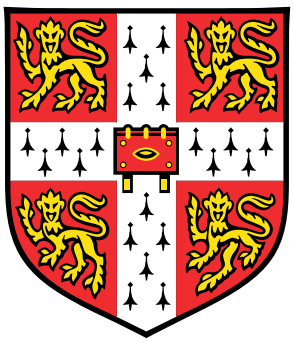
Optimal Importance Sampling in Quantum Monte Carlo for Lattice Models



Blaž Stojanovič

Supervisor: Prof. A. Lamacraft

Department of Physics

This dissertation is submitted for the degree of
Master of Philosophy in Scientific Computing

St. John's College June 2021

I would like to dedicate this thesis to …

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 15,000 words including appendices, figure legends, and tables.

Blaž Stojanovič
June 2021

Acknowledgements

- Zala
- Parents/Family
- Nikos/Tutor/Sally
- Prosen/Kosec+Jure
- Austen

Abstract

This is where you write your abstract ...

Table of contents

List of figures

List of tables

Nomenclature

Other Symbols

$\mathbb{E}$	Expectation
"$\mathcal{F}$"	σ-field (algebra)
$\mathbb{F}$	Filtration
$\mathbb{P}$	Measure
$\mathcal{N}$	The Gaussian distribution
$\{X\}$	Random variable
$\mathbb{R}$	The set of real numbers
$\{X_t\}$	Stochastic process
Cov	Covariance
Var	Variance
W_t	Wiener process, mathematical Brownian motion

Acronyms / Abbreviations

cdf	Cumulative density function
CNN	Convolutional Neural Network
DL	Deep Learning
DMC	Diffusion Quantum Monte Carlo
e.g.	Exempli gratia ("for the sake of an example")

FP	Fokker-Planck
GAN	General Adversarial Network
i.e.	Id est ("it is")
i.i.d	Independent and identically distributed
ML	Machine Learning
NN	Neural Network
p.b.c	Periodic boundary condition
pdf	Probability density function
QMC	Quantum Monte Carlo
SDE	Stochastic Differential Equations
s.p.	Stochastic process
s.t.	Such that
VAE	Variational Autoencoder
VMC	Variational Quantum Monte Carlo
w.r.t	With respect to

Chapter 1

Introduction

1.1 Thesis Contributions

1.2 Thesis Structure

Chapter 2

On the quantum many-body problem

2.1 Lattice models

2.1.1 Historical introduction

- I think starting from the magnetism point of view might be the best way to go, slowly lead into the field of condensed matter theory and lattice models.

2.1.2 The Schrödinger equation and the Feynman path integral

The wavefunction

$$\Psi(r_1, \ldots, r_N) \tag{2.1}$$

the Schrödinger equation

$$i\frac{\partial \psi(\boldsymbol{r}, t)}{\partial t} = \hat{H}\psi(\mathbf{r}, t) \tag{2.2}$$

for a single particle in an external potential $\hat{V}(\boldsymbol{r})$ the Hamiltonian is

$$\hat{H}\phi(\boldsymbol{r}) = -\frac{1}{2}\nabla^2\phi(\boldsymbol{r}) + \hat{V}(\boldsymbol{r})\phi(\boldsymbol{r}). \tag{2.3}$$

Alternatively to the Schrödinger equation one can use an integral Green's function representation to express the wavefunction ψ at some future time t_2 given initial condition $\psi(\boldsymbol{r}, t_1)$ as

$$\psi(\boldsymbol{r}_2, t_2) = \int \mathcal{K}(\boldsymbol{r}_2, t_2; \boldsymbol{r}_1, t_1)\, \psi(\boldsymbol{r}_1, t_1)\, \mathrm{d}\boldsymbol{r}_1. \tag{2.4}$$

The solution to equation

$$\left(i\frac{\partial}{\partial t_2} - H_{\boldsymbol{r}_2}\right)\mathcal{K}(\boldsymbol{r}_2, t_2; \boldsymbol{r}_1, t_1) = i\delta(\boldsymbol{r}_1 - \boldsymbol{r}_2)\,\delta(t_1 - t_2) \tag{2.5}$$

and the *propagator* $\mathcal{K}(\boldsymbol{r}_2, t_2; \boldsymbol{r}_1, t_1)$ is expressed using the Feynman path integral

$$\mathcal{K}(\boldsymbol{r}_2, t_2; \boldsymbol{r}_1, t_1) = \int_{\substack{\boldsymbol{r}(t_1)=r_1 \\ \boldsymbol{r}(t_2)=r_2}} \mathcal{D}\boldsymbol{r}(t) \exp\left(i\int_{t_1}^{t_2} \mathcal{L}(\boldsymbol{r}, \dot{\boldsymbol{r}})dt\right), \tag{2.6}$$

where $\mathcal{L}$ is the classical Lagrangian function of the system

$$L(\boldsymbol{r}, \dot{\boldsymbol{r}}) = \frac{1}{2}\dot{\boldsymbol{r}}^2 - \hat{V}(\boldsymbol{r}), \tag{2.7}$$

and the integral is over all paths that satisfy the endpoint conditions.

2.1.3 Examples of lattice models

$$\hat{\sigma}_i^x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}_i \quad \hat{\sigma}_i^y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}_i \quad \hat{\sigma}_i^z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}_i \tag{2.8}$$

Transverse-field Field Ising model

$$\hat{H}_{\text{Ising}} = -J\sum_{\langle i,j\rangle} \hat{\sigma}_i^z \hat{\sigma}_j^z - h\sum_i \sigma_i^x \tag{2.9}$$

Heisenberg model

$$\hat{H}_{\text{Heisenberg}} = -\frac{1}{2}\sum_{j=1}^{N}\left[J_x\hat{\sigma}_j^x\hat{\sigma}_{j+1}^x + J_y\hat{\sigma}_j^y\sigma_{j+1}^y + J_z\hat{\sigma}_j^z\hat{\sigma}_{j+1}^z + h\hat{\sigma}_j^z\right] \tag{2.10}$$

Bose-Hubbard model

$$\hat{H}_{\text{BH}} = -t\sum_{\langle i,j\rangle} \hat{b}_i^\dagger \hat{b}_j + \frac{U}{2}\sum_i \hat{n}_i(\hat{n}_i - 1) - \mu\sum_i \hat{n}_i \tag{2.11}$$

2.2 Solutions to the many-body problem and QMC

- Mention Complexity of QMC for fermions and bosons
- A more broad overview of methods that can be used for studying the many body problem (DRMG), with particular applications to lattice methods
- Partition QMC into nicer subsections

The Schrödinger equation underpins a large part of quantum chemistry and solid state physics. However, the quantum many-body problem, which amounts to solving the $3N$-dimensional Schrödinger equation is notoriously hard to solve. Ever since the postulation of the equation in 1925, great efforts have been made in solving the equation, both analytically and numerically. Perhaps most impactful was the development of various approximate methods to solve the many-body problem with the available computational resources. Hartree-Fock (HF) approaches solve an auxiliary system of independent electrons in a self-consistent field and assume that the wave function (for fermions) can be represented as a single Slater determinant. The HF method does not include electron correlation, which makes it a good approximation only in systems where correlation contributions are small. Post-HF methods, such as Coupled Cluster, Configuration interaction and Møller-Plesset theory include correlation by considering a linear combination of determinants. They can be extremely accurate but come at a high computational cost.

One of the most popular approaches used today is Density Functional Theory (DFT). It reformulates the many-body electron problem in terms of the 3-dimensional electron density $n(\mathbf{r})$, which is found by minimising the total energy functional $E[n(\mathbf{r})]$ [15]. DFT provides an alternative line of thought to the truncated Hilbert space of single particle orbitals [22] and is used extensively for simulating large systems as linear scaling variants of DFT exist [37]. While DFT is theoretically exact the true energy functional $E[n(\mathbf{r})]$ is not known and its parameterisations employ more accurate *ab initio* methods. One of which being Quantum Monte Carlo (QMC).

2.2.1 Quantum Monte Carlo

Quantum Monte Carlo is a class of methods that uses statistical sampling to directly deal with high-dimensional integration that arises from working with the many-body wave function. QMC methods are among the most accurate achieving chemical accuracy for smaller systems [11], and can achieve any degree of statistical precision sought. Quantum Monte Carlo is also very versatile and can be applied at both zero and finite temperatures [3]. The most

basic zero temperature QMC method is variational QMC (VMC). The method is composed of two parts, firstly it directly evaluates the variational energy $E_V = \langle \Psi_T | \hat{H} | \Psi_T \rangle / \langle \Psi_T | \Psi_T \rangle$ of the system using Monte Carlo integration and a trial wave function Ψ_T. Secondly the parameters of the trial wave function are optimised such as to minimise the variational energy E_V, giving the method its name. The first application of VMC was to ground state ^{4}He [24] and was later extended for studying many-body fermionic systems [6]. A way of obtaining excitation energies using VMC is to use a trial wave function that models an excited state of the system, if the trial wave function obeys a certain symmetry, the variational principle guarantees that this VMC energy calculation gives an upper bound on the lowest exact eigenstate of this symmetry. Furthermore, the method can be extended to study non-equilibrium properties of bosonic [5, 4], and fermionic [16] systems. The main advantage of VMC is its simplicity while the main drawback is that the accuracy is limited by the flexibility and form of the trial wave function [3]. As such VMC is usually employed as a first step in more advanced QMC simulations.

Projector quantum Monte Carlo (PMC) is a class of QMC methods which are in essence nothing more than stochastic implementations of the power method to obtain the dominant eigenvector of a matrix or a kernel function [12]. Their distinct advantage over VMC is that they are not constrained by our parametrisation of the trial wave function, as they can describe arbitrary probability distributions. The projector $\hat{P}$ has to be chosen in such a way, that the ground state of the system becomes the dominant eigenvector, i.e. $|\Psi_0\rangle = \lim_{n\to\infty} \hat{P}^n |\Psi_T\rangle$. Different ways of achieving this give rise to different flavours of PMC methods, e.g choice of space (real or orbital space) in which the walk is done and choosing either first or second quantisation. Using an exponential projector $\hat{P} = e^{\tau(E_T \mathbb{1} - \hat{H})}$ can be interpreted as propagation in imaginary time $\tau \longrightarrow it$ in turn transforming the Schrödinger equation into a diffusion equation, which is a continuous limit of the random walk and lends itself to stochastic integration [31]. Directly sampling from the exact Green function is known as Projector Green Function Monte Carlo (GFMC) method [18, 19]. A convenient approximation to GFMC is its short-time approximation which leads to one of the most popular QMC methods, diffusion Monte Carlo (DMC) [11, 31]. In this regime one can exploit analytical solutions to diffusion and rate problems to write an explicit form of the Green's function. Additionally, by using the Trotter-Suzuki formula the time-step bias can be expressed and accounted for [3]. DMC is statistically implemented by using a population of walkers which either branch or die, over which the average is calculated. Reptation quantum Monte Carlo [31] (RMC) is an alternative formulation which only uses a single walker, and instead of branching and dying the MC moves mutate the path of that single walker. The use of a guiding wave function for importance sampling greatly

improves the statistical efficiency of PMC methods. The guiding wave function is usually obtained by means of VMC or some mean field calculation.

PMC methods suffer from the *sign problem*, which is present in Markov chain simulation of distributions that are not strictly positive, thus in fermionic and frustrated systems [12]. The problem refers to an exponential decrease in sampling efficiency with system size. The search for solutions of this problem is still an area of active research [11] but is in practice remedied by the *fixed-node* approximation [2]. It imposes a boundary condition into the projection, such that the projected state shares the same zero crossings (nodal surface) with a trial wave function, which is again usually obtained with VMC. The projected state is now only exact when the nodal surface is exact. Nevertheless this approximation is quite accurate [11]. Fixed node is widely used, one of its first applications being the electron gas [8], which is used in parameterisations of the exchange correlation functional in LSDA [38].

Quantum Monte Carlo methods have had a lot of success at finite temperatures. Auxiliary-field Monte Carlo, or Path Integral Monte Carlo [7], which leads to ring-polymer molecular dynamics, may be used for this purpose. Additionally QMC is not limited to continuum space applications and has been extensively used to study lattice models, notable examples being the cluster/loop algorithm and the worm algorithm [12, 29].

Quantum Monte Carlo methods are generally more computationally expensive than DFT approaches, but on the other hand QMC codes are, as a rule of thumb, simpler to implement. Furthermore, since the wave function does not need to be stored directly, QMC has reasonable storage requirements. The high computational cost of the QMC methods is remedied by the fact that they are intrinsically parallelisable, the core calculation involves generating (pseudo)-random numbers, performing a simple calculation and in the end averaging over the results. Therefore, implementations of QMC algorithms that have been applied to practical problems are optimised to run on massively parallel hardware with little overhead [26]. Finally, the repetitive nature of the Monte Carlo calculation lends itself to hardware acceleration using either graphical processing units (GPUs) or field-programmable gate arrays (FPGAs) [3].

Chapter 3

Feynman-Kac: connecting Quantum Mechanics and Stochastic Processes

In this chapter we will provide a bridge between the quantum many-body problem discussed in the previous chapter and stochastic processes. This will entail introducing the Feynman-Kac formula and relating it to the Fokker-Planck equation and optimal control approaches to QM. Moreover, a probabilistic view of the cost function will lead us to proposals for loss functions that can be used to learn optimal transition rates and consequently sample the ground state.

The field of stochastic processes is a vast body of work, approached from different angles by mathematicians, physicists and engineers. A necessary consequence of this is that the literature ranges from extremely thorough and rigorous [32, 33] to more applied and intuitive [35]. For this reason, the mentioned discussion will be preceded by an overview of the mathematical notation, lemmas and results from stochastic processes and measure theory that underpin some core ideas of this thesis. To avoid including a whole textbook of material on measure and stochastic processes some concepts will not be rigorously defined, the text will point to relevant literature where this is the case.

3.1 Stochastic processes

3.1.1 Fundamentals

This brief, more formal, discussion of stochastic processes is based mostly upon classic texts [10, 32, 33] and borrows some intuitions from [35]. The most basic quantity that we will need is the **probability space**.

Definition 3.1.1 (Probability space) *The probability space is a tuple* $(\Omega, \mathcal{F}, \mathbb{P})$*, where* Ω *is the* ***sample space****,* $\mathcal{F}$ *is a* σ*-****field****, and* $\mathbb{P}$ *is the* ***measure****.*

The sample space is simply the set of all possible outcomes. A canonical example would be the roll of a 6-sided dice, $\Omega = \{1, 2, 3, 4, 5, 6\}$. Without measure $\mathbb{P}$, the tuple $(\Omega, \mathcal{F})$ is termed a **measurable space**.

Definition 3.1.2 (σ-field) *A* σ*-field* $\mathcal{F}$ *on a set* Ω*, is a nonempty collection of subsets of* Ω *that includes* Ω *itself, is closed under complement, i.e. if* $A \in \mathcal{F}$ *then* $A^c \in \mathcal{F}$*, and is closed under countable unions,* $\cup_i A_i \in \mathcal{F}$ *if* $A_i \in \mathcal{F}$ *is a countable union of sets.*

The main utility of the σ-field to us is its use in defining measures. We want to be able to assign a non-negative real number to all subsets of Ω, as well as the size of the union of the disjoint sets to be the sum of their individual sizes. This is not always possible, a counterexample for the real line being Vitali sets. The collection $\mathcal{F}$, must thus only include *measurable* sets, which are precisely the ones that satisfy the constraints imposed by the σ-field.

Definition 3.1.3 (Measure) *A non-negative countably additive set function* $\mu : \mathcal{F} \longrightarrow \mathbb{R}$ *that satisfies*

i) $\mu(A) \geq \mu(\emptyset) = 0$ *for all* $A \in \mathcal{F}$

ii) if $A_i \in \mathcal{F}$ *is a countable sequence of disjoint sets, then* $\mu(\cup_i A_i) = \sum_i \mu(A_i)$

is a ***measure****.*

If $\mu(\Omega) = 1$, then μ is a **probability measure** and will be denoted by $\mathbb{P}$. With this notion we are now able to define a random variable (r.v) and a stochastic process (s.p.).

Definition 3.1.4 (Random variable) *A* ***random variable*** X *defined on* Ω *is a real-valued measurable function* $X(\omega)$*,* $X : \Omega \longrightarrow \mathbb{R}^d$*.*

For a function to be measurable, we require that its preimage X^{-1} is in the σ-field $\mathcal{F}$

$$X^{-1}(B) = \{\omega : X(\omega) \in B\} \in \mathcal{F}, \tag{3.1}$$

and that this holds for every Borel set B in the Borel σ-field [1] of $\mathbb{R}^d$, which is simply the smallest σ-field that contains all measurable sets in $\mathbb{R}^d$.

[1] For a proper definition of the Borell set see ch. 3 of [34].

A random variable X induces a probability measure μ on $\mathbb{R}^d$ called its **distribution**, this is done by setting $\mu(A) = P(X \in A)$ for Borel sets A. Moreover, the distribution is usually given in terms of a **distribution function** $F(x)$

$$F(x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}) = \mathbb{P}(X \leq x), \tag{3.2}$$

and X is said to have a **density function** $f(x)$ if $F(x)$ can be written as

$$F(x) = \int_{-\infty}^{x} f(y)\mathrm{d}y. \tag{3.3}$$

In essence, the random variable provides a connection between the less familiar probability measure $\mathbb{P}$ and the cumulative distribution function (CDF).

3.1.2 Stochastic process

Definition 3.1.5 (Stochastic process) *Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a measurable (state) space $(E, \mathcal{E})$, we define the collection $\{X_t : t \in T\}$ of set T indexed and $(E, \mathcal{E})$ valued random variables a **stochastic process**.*

By far the most common case for the index set T, is time $T = \mathbb{R}^+$. Such s.p's are called *temporal*, examples include the model of velocity of a Brownian particle under influence of friction, in Fig. 3.1, or the Black-Scholes model. Nevertheless, the index set is not limited to time, as is often the case with Gaussian Process regression [30]. In this thesis we will mostly deal with temporal s.p's of the kind that do not "see into the future". This notion is formalized using **filtrations**. A filtration $\mathbb{F} = (\mathcal{F}_t)_{t \in T}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is just an increasing sequence or order of σ-fields

$$\mathcal{F}_s \subset \mathcal{F}_t \text{ if } 0 \leq s \leq t < \infty \tag{3.4}$$

The filtration associated to a process that records its "past behaviour" at each time is called the **natural filtration**.

Definition 3.1.6 (Adapted process) *A process $\{X_t\}$ is said to be **adapted to the filtration** $(\mathcal{F}_t)_{t \in T}$ if the random variable $X_t : \Omega \longrightarrow E$ is $\mathcal{F}_t$-measurable function for each $t \in T$.*

A process that is *non-anticipating*, i.e. depends only on the past and present, is adapted to the filtration $(\mathcal{F}_t)_{t \in T}$.

Definition 3.1.7 (Brownian motion) ***Brownian motion** or a non-anticipating **Wiener process** is a stochastic process W_t, with the following properties:*

i) $W_0 = 0$

ii) W_t *is almost surely continuous in* t

iii) W_t *has independent increments*

iv) $W_t - W_s \sim \mathcal{N}(0, t-s)$ *for* $0 \leq s \leq t$

A realisation of Brownian motion can be found in Fig. 3.1.

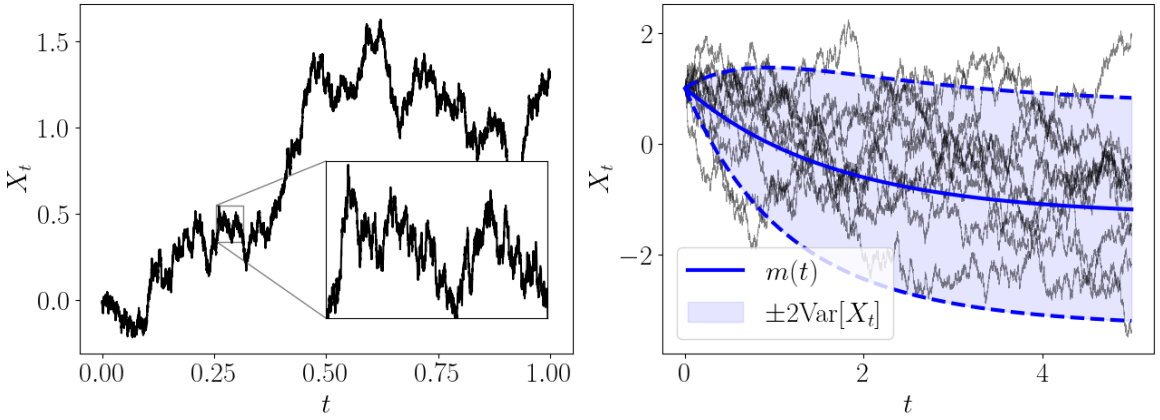


Fig. 3.1 **Brownian motion and Ornstein–Uhlenbeck process. left:** A single realisation of the Brownian process. **right:** Mean, variance and 10 samples of the Ohrnstein-Uhlenbeck process with $\theta = 0.6, \sigma = 1.1, X_0 = 1.0, \mu = -1.3$, integrated using Euler-Maruyama method.

3.1.3 Integrals

In order to proceed and define stochastic differential equations (SDE's) and the Radon-Nikodyn derivative, we must spend some time discussing various integrals we will use. In particular, alongside the usual Riemann integral, we will need three more types of integrals, which we will briefly describe without mathematical derivation. The simplest kind of integral we will introduce is the integral of a stochastic process

$$I = \int_0^t X_t \mathrm{d}t. \tag{3.5}$$

The simple appearance of the integral is deceiving as the integrand is a realisation of a $\mathcal{F}_t$-adapted stochastic process $\{X_t\} : \Omega \times T \longrightarrow \mathbb{R}^d$, meaning that I itself is a random variable. However, since each realisation of X_t is almost surely continuous, I can be expanded as a Riemann sum, which converges under mean-squared norm to I, so long as the mean $\mathbf{m}(t) = \mathbb{E}[X_t]$ and covariance $\mathbf{k}(t,s) = \mathrm{Cov}(X_t, X_s)$ are continuous. In practice, computing the

mean and covariance of I is usually enough to understand the resulting stochastic process. Importantly, integrals of continuous functions of s.p's $h(X_t)$, $h : \mathbb{R} \longrightarrow \mathbb{R}$ can be computed in a similar manner.

The second type of integrals we need to consider, are integrals with respect to a s.p, known as **Itô integrals**

$$Y_t = \int_0^t H_s \mathrm{d}X_s, \tag{3.6}$$

where both H_s and X_s are stochastic processes. The result integral Y_t is itself a stochastic process which resides in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, filtered by $(\mathcal{F}_t)_{t\in T}$. The integral can be formalised by putting slight constraints on what sort of stochastic processes X_s and H_t can be, expanding Y_t as a Riemann sum and proving convergence. Details of this procedure can be found in [33].

Finally we must define the **Lebesgue-Stieltjes integral** [13], which we need to properly define expectations of stochastic processes.

Definition 3.1.8 (Lebesgue-Stieltjes Integral) *Given probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *and measurable function* $f : \Omega \longrightarrow \mathbb{R}$*, the* ***Lebesgue-Stieltjes integral***

$$I = \int_A f(x) \mathrm{d}\mathbb{P}(x), \tag{3.7}$$

is the Lebesgue integral[2] *with respect to measure* $\mathbb{P}$*,* $A \in \mathcal{F}$*.*

With it we can define expectations in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ as

$$\underset{\mathbb{P}}{\mathbb{E}}[f(x)] = \int_\Omega f(x) d\mathbb{P}(\boldsymbol{x}). \tag{3.8}$$

For a newcomer to stochastic processes this formulation may seem redundant, can we not just calculate expectations using a Riemann integral and the PDF? We can, and when the distribution $\mathbb{P}$ can be expressed in terms of the PDF (3.3), the Lebesgue integral can be interpreted in this way. However, stochastic processes need not admit a PDF, that is when the Lebesgue-Stieltjes integral is necessary.

3.1.4 Stochastic Differential Equations

In this thesis we will refer to a SDE as an informal notation of an Itô integral equation or **Itô process**.

[2] For proper definition of the Lebesgue integral see ch. 1 of [34].

Definition 3.1.9 (Itô process) *Given deterministic functions* $v : \mathbb{R}^d \times \mathbb{R}^+ \longrightarrow \mathbb{R}^d$ *and* $\sigma : \mathbb{R}^d \times \mathbb{R}^+ \longrightarrow \mathbb{R}^{d\times d}$*, we define the* ***Itô process*** X_t *as the sum of Itô and Lebesgue integrals*

$$X_{t+s} - X_t = \int_t^{t+s} \sigma(X_u, u)\,\mathrm{d}W_u + \int_t^{t+s} v(X_u, u)\,\mathrm{d}u, \tag{3.9}$$

where W_t *is a Brownian motion.*

In simplified notation we can write (3.9) as

$$\mathrm{d}X_t = \sigma(X_t, t)\,\mathrm{d}W_t + v(X_t, t)\,\mathrm{d}t, \tag{3.10}$$

this is what we refer to as an SDE, an example can be found in Fig. 3.1. The functions v and σ, we will refer to as the **drift** and **volatility** of the Itô process respectively. The most intuitive interpretation of a SDE is in terms of the time evolution of the PDF of the process X_t. It is described by the **Fokker-Planck equation** [3]

$$\frac{\partial p(\mathbf{x}, t)}{\partial t} = -\sum_{i=1}^{N} \frac{\partial}{\partial x_i}\left[\mu_i(\mathbf{x}, t)p(\mathbf{x}, t)\right] + \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\partial^2}{\partial x_i \partial x_j}\left[D_{ij}(\mathbf{x}, t)p(\mathbf{x}, t)\right], \tag{3.11}$$

where $p(\mathbb{x}, t)$ is the PDF of the solution to the SDE and $\mathbf{D} = \frac{1}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^\top$ is the diffusion tensor. Finally we state without proof a consequence of Itô calculus, most commonly named **Itô's rule** or **lemma**, it is the stochastic calculus equivalent of the chain rule

Lemma 3.1.1 (Itô's lemma) *Given an Itô process* X_t *as given by* (3.9) *and a twice differentiable scalar function* $f(X_t, t)$*, then the Itô process for* f *is*

$$\mathrm{d}f = \frac{\partial f}{\partial t}\mathrm{d}t + \sum_i \frac{\partial f}{\partial x_i}\mathrm{d}x_i + \frac{1}{2}\sum_{ij} \frac{\partial^2}{\partial x_i \partial x_j} f\,\mathrm{d}x_i \mathrm{d}x_j, \tag{3.12}$$

when compared to ordinary calculus we notice an additional quadratic term.

3.1.5 Radon-Nikodym Derivative and Girsanov theorem

To perform importance sampling we perform a change of measure in an integral

$$\int_A f(x)\mathrm{d}\mathbb{P}(x) = \int_A f(x)\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x)\mathrm{d}\mathbb{Q}(x). \tag{3.13}$$

[3] Derivation in [35].

The function that measures the rate of change of density of one measure w.r.t another is the **Radon-Nikodym derivative** $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x)$.

Theorem 3.1.2 (Radon-Nikodym theorem) *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on the measurable space $(\Omega, \mathcal{F})$, then the measurable function **Radon-Nikodym derivative** $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x) : \Omega \longrightarrow [0, \infty)$ exists and*

$$\mathbb{P}(A) = \int_A \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x)\mathrm{d}\mathbb{Q}(x), \tag{3.14}$$

for set $A \subseteq \mathcal{F}$.

The RN derivative will also be useful in defining the KL divergence between two **path measures**. Properly defining the path measure would bring a lot of notational overhead, it is enough to think of it as a measure on the **path space**, i.e all possible paths of a SDE, for rigour see [23]. Finally, we state the **Girsanov theorem** that is often used for transforming or removing drift functions of SDE, it is the RN derivative between an Itô process and one with $v = 0$ and $\sigma = 1$, i.e. Brownian motion.

Theorem 3.1.3 (Girsanov Theorem) *Given Itô process*

$$\mathrm{d}X_t = \mathrm{d}W_t + v(X_t, t)\mathrm{d}t \quad \textit{and} \quad X_0 = 0 \tag{3.15}$$

and Brownian motion $\mathrm{d}Y_t = \mathrm{d}W_t$, the RN derivative of their respective path measures $\mathbb{P}$ and $\mathbb{P}_0$ is

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}_0} = \exp\left(-\frac{1}{2}\int_0^t |v(X_s, s)|^2\,\mathrm{d}s + \int_0^t v(X_s, s)^\top \mathrm{d}W_s\right) \tag{3.16}$$

This *change in dynamics* as we will call it later is true in the sense, that expectations for an arbitrary functional $h(\cdot)$ of the path from 0 to t are

$$\mathop{\mathbb{E}}_{\mathbb{P}_0}[h(X_t)] = \mathop{\mathbb{E}}_{\mathbb{P}_0}\left[\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}_0}h(Y_t)\right]. \tag{3.17}$$

For a more general case and proof see [35].

3.1.6 Markov processes

We now shift our view to a special kind of s.p's, ones that satisfy the **Markov property** called **Markov processes** or **Markovian**. The property is sometimes referred to as *memorylessness*, as the future of a Markov process depends only on the present state. We can classify the processes based on the system's **state-space** S, which can be either discrete (countable) or continuous, and the **time indexing** of the system, either discrete-time

$\{X_n\}_{n\geq 0}$ or continuous-time $\{X_t\}_{t\geq 0}$. A taxonomy is given in Table 3.1. We will not specifically discuss Markov processes in continuous state-space, but it is important to note that any Itô process with time-homogenous drift $v = v(X_t)$ and volatility $\sigma = \sigma(X_t)$ is Markovian.

From now on we refer to Markov processes in countable state-space as **Markov chains**. We base our discussion on [32] and [28].

Table 3.1 **Taxonomy of Markov processes**

	Countable state-space	Continuous state-space
Discrete time	**index:** $\{X_n\}_{n\geq 0}$, $n \in \mathbb{Z}^+$ **state-space:** countable set I **define:** stochastic $\{P\}_{ij}$ **example:** DTMC	**index:** $\{X_n\}_{n\geq 0}$, $n \in \mathbb{Z}^+$ **state-space:** general state-space Ω **define:** stochastic kernel K **example:** Harris Chain
Continuous time	**index:** $\{X_t\}_{t\geq 0}$, $t \in \mathbb{R}^+ = [0,\infty)$ **state-space:** countable set I **define:** rate $\{\Gamma\}_{ij}$ equiv. to jump chain $\{J_n\}_{n\geq 0}$ and hold times $\{S_n\}_{n\geq 1}$. **example:** CTMC	**index:** $\{X_t\}_{t\geq 0}$, $t \in \mathbb{R}^+ = [0,\infty)$ **state-space** general state-space Ω **define:** stochastic kernel K **example:** Diffuson process

Discrete-time Markov Chains

The simplest and most common Markov process is a Markov chain in discrete time, an example of it can be found in Fig. 3.2. Its state-space is a countable set I and we call each $i \in I$ a **state**. We define a distribution λ in a familiar way

$$\lambda = \{\lambda_i : i \in I\} \quad \text{where} \quad \forall i : \ 0 \leq \lambda_i < \infty \quad \text{and} \quad \sum_{i\in I} \lambda_i = 1. \tag{3.18}$$

We can now set λ as a distribution of some random variable $X : \Omega \longrightarrow I$ as

$$\lambda_i = \mathbb{P}(X = i) = \mathbb{P}(\{\omega : X(\omega) = i\}), \tag{3.19}$$

where we are still working in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A discrete-time Markov chain is defined in terms of its **transition matrix** $P = \{p_{ij} : i,j \in I\}$, which is a **stochastic matrix** meaning all of its rows $\{p_{ij} : j \in I\}$ are distributions.

Definition 3.1.10 (Discrete-time Markov chain) *A discrete time stochastic process* $\{X_n\}_{n\geq 0}$ *is a* ***discrete-time Markov chain*** *with initial distribution* λ *and transition matrix* P *if for* $i_1, \ldots, i_n + 1 \in I$ *and* $n \geq 0$

i) $\mathbb{P}(X_0 = i_1) = \lambda_{i_1}$

ii) $\mathbb{P}(X_{n+1} = i_{n+1} \mid X_0 = i_0, \ldots, X_n = i_n) = p_{i_n i_{n+1}}$

Rewriting the second condition above, it is clear that the Markov chain is without memory

$$\mathbb{P}(X_{n+1} = i_{n+1} \mid X_0 = i_1, \ldots, X_n = i_n) = \mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n). \tag{3.20}$$

Intuitively we understand the discrete-time Markov chain as a system changing its state at discrete time intervals, each time choosing the next state according to the row of the Transition matrix corresponding to the current state.

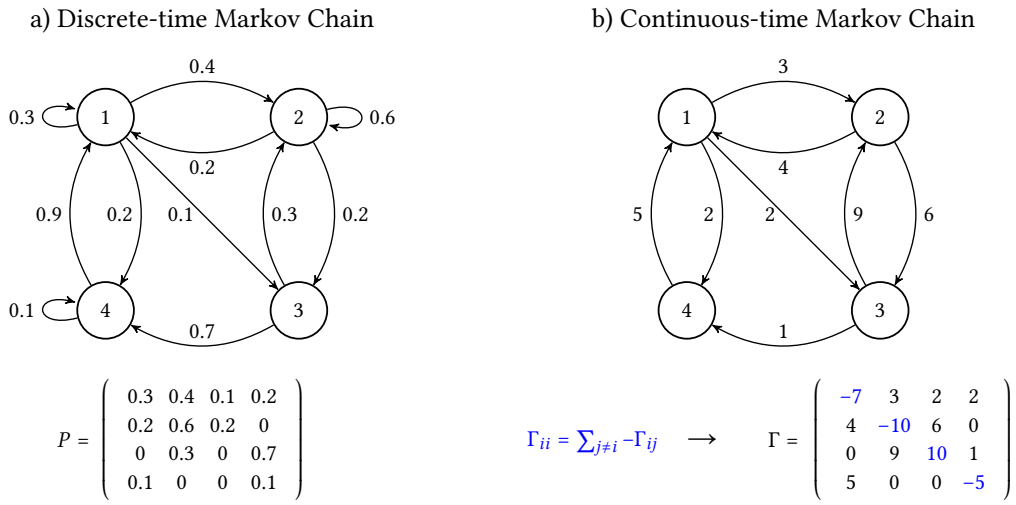


Fig. 3.2 **Discrete and continuous time Markov Chains. left:** Discrete-time Markov Chain defined by P. **right:** Continuous-time Markov Chain defined by Γ.

Continuous-time Markov Chains

Defining a Markov chain in continuous time is slightly trickier as describing the system with a stochastic matrix does no longer suffice because transition probabilities become zero when considering an infinitesimal time. Instead a continuous-time Markov Chain (CTMC) is characterised by a **rate matrix** or **infinitesimal generator matrix** Γ defined on the set I. A rate matrix has the following three properties

i) $0 \leq \Gamma_{ii} < \infty, \quad \forall i$

ii) $\Gamma_{ij} \geq 0, \quad \forall i \neq j$

iii) $\sum_{j \in I} \Gamma_{ij} = 0, \quad \forall i$

While the CTMC can be interpreted in a number of ways, we shall use the so called **jump chain** and **holding times** representation, see also Fig. 3.3. We can think of a CTMC as a

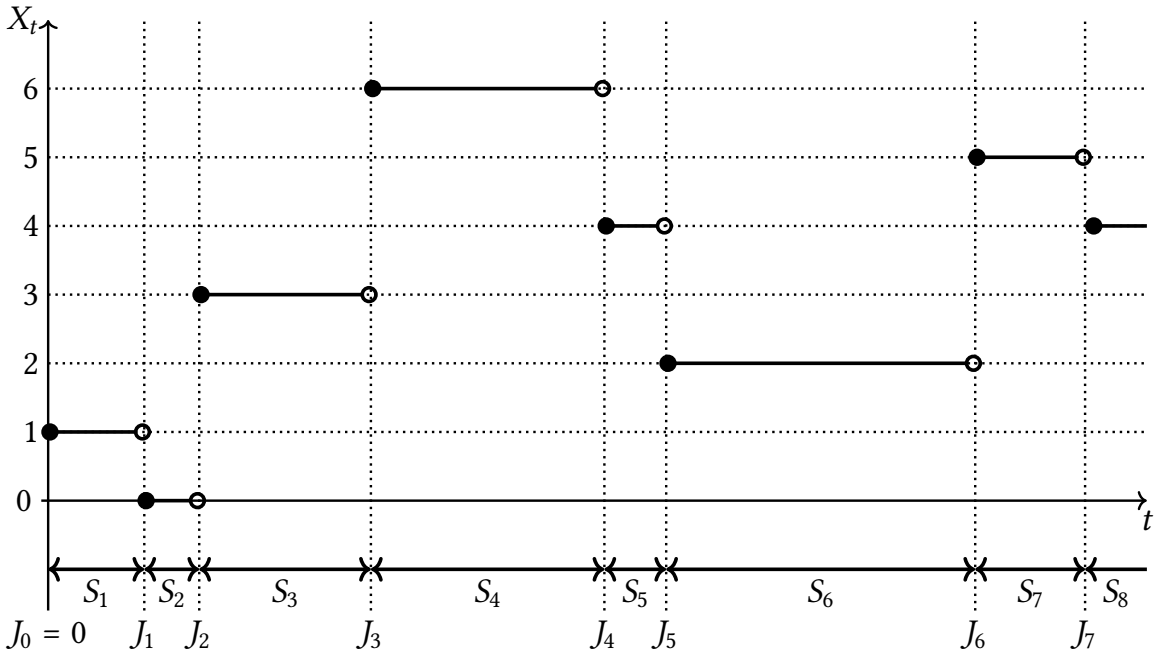


Fig. 3.3 **Jump chain and Holding times.** A discrete space Markov process $\{X_t\}_{t \geq 0}$ in continuous time. The holding times S_n are independent exponential random variables and the transition probabilities at jump times J_n are given with the jump matrix Π. Inspired by [28].

series of discrete jumps, where the system remains in each state for a certain holding time. This suggests that we can construct the CTMC from a discrete-time chain with stochastic matrix Π, which we will call the **jump matrix**, and a set of independent random variables $\{S_n\}$ which determine the holding times. We construct matrix Π by rescaling rows of Γ so they add up to one, putting a 0 on the diagonal

$$\Pi_{ij} = \begin{cases} \Gamma_{ij}/\Gamma_{ii} & \text{if } j \neq i \text{ and } \Gamma_{ii} \neq 0 \\ 0 & \text{if } j \neq i \text{ and } \Gamma_{ii} = 0 \end{cases}$$
$$\Pi_{ii} = \begin{cases} 0 & \text{if } \Gamma_{ii} \neq 0 \\ 1 & \text{if } \Gamma_{ii} = 0. \end{cases} \tag{3.21}$$

In order for the process to possess the Markov property, the distribution of holding times $\{S_n\}$ must be exponential [28],

$$S_{n+1} \sim \text{Exp}(-\Gamma_{ii}(X_n)), \tag{3.22}$$

with exponential parameters being $-\Gamma_{ii}$ where i is the current state. Processes with different holding time distributions are called **semi-Markov**. The jump times $\{J_n\}$ are simply

$$J_n = S_1 + \ldots + S_n. \tag{3.23}$$

Definition 3.1.11 (Continuous-time Markov chain) *A stochastic process $\{X_t\}_{t\geq 0}$ on set I is a **continuous-time Markov chain** if its jump chain $\{Y_n\}_{n\geq 0}$ is a discrete-time Markov chain and its holding times $\{S_n\}_{n\geq 1}$ are independent exponential random variables $S_n \sim \text{Exp}(-\Gamma_{ii}(X_n))$.*

An equivalent formulation is in terms of **competing exponentials**. Transitions $\Gamma_{j\to k}$ from j to k are defined as independent exponential random variables $\tau_{j\to k}$

$$\tau_{j\to k} \sim \text{Exp}(\Gamma_{jk}), \quad j \neq k \tag{3.24}$$

the next state is then chosen as

$$Y_{n+1} = \text{argmin}_k \, \tau_{j\to k}. \tag{3.25}$$

The chain $\{Y_n\}_{n\geq 0}$ along with times

$$S_n = \min_k \tau_{j\to k}, \tag{3.26}$$

gives the full description of the CTMC. With this formulation in mind we now interpret Γ_{ii} as the rate ob *leaving* current state and Γ_{ij} as the rate of *going* from i to j.

3.2 The Feynman-Kac formula

The Feynman path integral formulation (2.6) was extensively used by physicists for decades, even in the absence of a formal mathematical formulation which is hard to define because of the difficulties with defining an appropriate measure on the path space. Kac [17] provided a rigorous formulation of the *real-valued* case of the Feynman path integral, and the resulting

Feynman-Kac formula provides a bridge between *parabolic* partial differential equations and stochastic processes.

To illustrate the Feynman-Kac formula let us consider a single particle with Hamiltonian

$$\hat{H} = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} + V(x) \tag{3.27}$$

and the Schrödinger equation in *imaginary time*, which is of the elliptic type,

$$\partial_t |\psi_t\rangle = -\hat{H}|\psi_t\rangle. \tag{3.28}$$

Its formal solution, the time propagation of an initial wave function $|\phi_0\rangle$ at $t = 0$, is written as

$$|\psi_t\rangle = e^{-\hat{H}t}|\psi_0\rangle. \tag{3.29}$$

From the spectral decomposition of the operator $e^{-\hat{H}t}$ in terms of eigenstates $|\phi_n\rangle$ and eigen-energies E_n of the Hamiltonian $\hat{H}$

$$e^{-\hat{H}t} = \sum_n e^{-E_n t}|\phi_n\rangle\langle\phi_n|, \tag{3.30}$$

it follows that the term corresponding to the ground state of the system $|\phi_0\rangle$ decays the slowest. Thus starting in some initial state and propagating for a long imaginary time it leads into the ground state with the decay rate giving the ground state energy as

$$\lim_{t\to\infty} |\psi_t\rangle \propto e^{-E_0 t}|\phi_0\rangle, \tag{3.31}$$

where E_0 is the ground state energy and $|\phi_0\rangle$ is the corresponding state. Kac noticed that the kinetic term of the Lagrangian in (2.6) could be interpreted as a measure on Brownian walks, and a solution to the imaginary time Schrödinger equation can be written as

$$\psi(x,t) = \mathop{\mathbb{E}}_{X\sim \text{ Brownian with } X_t=x}\left[\exp\left(-\int_0^t V(X_\tau,\tau)\,\mathrm{d}\tau\right)\psi(X_0,0)\right], \tag{3.32}$$

where only the **endpoint** at time t of the Brownian process fixed, whereas the starting point at time $t = 0$ is not, $\psi(x,0)$ encodes the initial condition into this representation. When there is no external potential $V(x) = 0$, the Schrödinger equation in imaginary time is the

diffusion equation and the Feynman-Kac solution is simply

$$\begin{aligned}\psi(x,t) &= \mathop{\mathbb{E}}_{X\sim \text{ Brownian with } X_t=x}\left[\psi\left(X_0,0\right)\right] \\ &= \frac{1}{\sqrt{2\pi t}}\int e^{-\left(x-x'\right)^2/2t}\psi_0\left(x'\right)\mathrm{d}x'\end{aligned} \tag{3.33}$$

An illustration of the Feynman-Kac approach to the problem with no external potential $V(x)$ in 1D is depicted in Fig. 3.4. This simple case, does not involve the potential $V(x)$. The role of the potential in the Feynman-Kac formula is to weight the Brownian paths, in turn defining the Feynman-Kac *path measure* $\mathbb{P}_{\text{FK}}$. A path measure is simply a measure on the path space and the Feynman-Kac measure is related to the Brownian measure $\mathbb{P}_0$ by the *Radon-Nykodym* derivative

$$\frac{\mathrm{d}\mathbb{P}_{\text{FK}}}{\mathrm{d}\mathbb{P}_0} = \mathcal{N}\exp\left(-\int V\left(X_t\right)dt\right), \tag{3.34}$$

where $\mathcal{N}$ is a normalizing constant. Intuitively we can understand the measure as assigning more weight to Brownian paths that spend more time in the attractive region ($V(x) < 0$) than in repulsive regions ($V(x) > 0$), this is illustrated in Fig. 3.5.

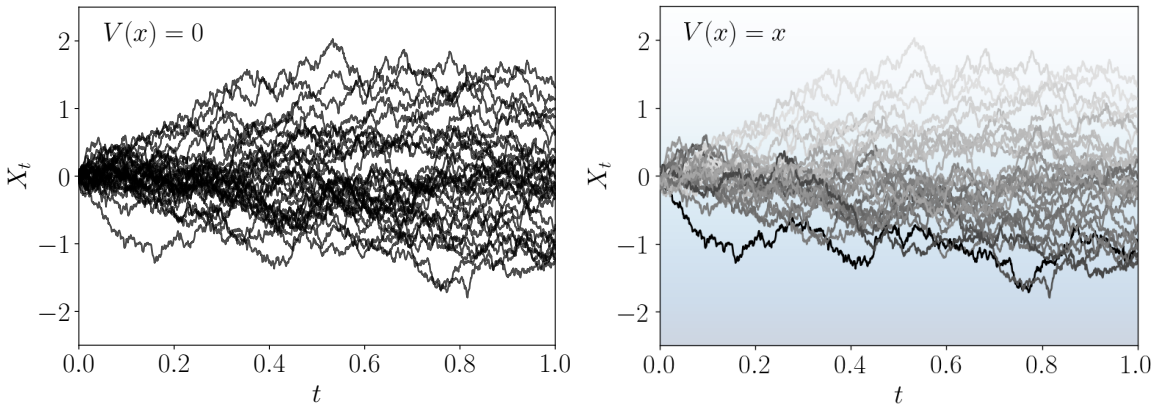


Fig. 3.5 **Feynman-Kac measure in a linear potential. left:** $N = 30$ Brownian paths. **right:** The paths colored by their likelihood under the Feynman-Kac measure with $V(x) = x$.

Moreover, this new stochastic process is Markovian. This is crucial for our approach because of the connection with SDEs. In the continuous case we have discussed so far, the mapping between the Fokker-Planck equation and the Schrödinger equation is straightforward in terms of a similarity transform. Starting from the FP equation with drift $v(x) = -U'(x)$

given as a gradient of some potential function $U(x)$ and its solution $\rho(t,x)$

$$\frac{\partial \rho}{\partial t} = \mathcal{L}_{\text{FP}}\rho = \frac{\partial}{\partial x}\left[\frac{\partial \rho}{\partial x} + U'(x)\rho\right], \tag{3.35}$$

we can define the function

$$\psi(x,t) = \frac{\rho(x,t)}{\sqrt{\rho_0(x)}}, \tag{3.36}$$

with ρ_0 being the stationary distribution of the FP equation

$$\frac{\partial}{\partial x}\left[\frac{\partial \rho}{\partial x} + U'(x)\rho\right] = 0 \quad \longrightarrow \quad \rho_0(x) \propto \exp(-U(x)). \tag{3.37}$$

Function $\psi(x,t)$ satisfies the imaginary time Schrödinger equation (3.28) with the Hamiltonian

$$H = -\frac{\partial^2}{\partial x^2} - \frac{U''}{2} + \frac{U'^2}{4}, \tag{3.38}$$

its ground state wavefunction is

$$\psi_0(x) = \sqrt{\rho_0(x)}. \tag{3.39}$$

In other words, the quantum ground state probability distribution $|\psi_0|^2$ is the same as classical stationary distribution ρ_0 of the stochastic process X_t

$$\mathrm{d}X_t = \mathrm{d}W_t + v(X_t)\,\mathrm{d}t, \tag{3.40}$$

in the literature referred to as the *Nelson's ground state process* [27, 1]. This connection serves as the backbone of our solution approach (revisit), as the ability to efficiently sample from the stochastic process is equivalent to sampling from the ground state of the quantum system. Even though the connection is simple, it comes with a caveat. Starting from the Schrödinger equation one needs to find the drift $v(x)$ and while the connection with the potential of the Hamiltonian is clear-cut in 1D, this is not the case in many-body systems, i.e. the *inverse problem* of finding the stochastic process of a given Hamiltonian is difficult, and is one of the core problems approached in this thesis.

3.3 Stoquastic Hamiltonians and Lattice-model representations

A similar connection between stochastic processes and the ground state exists in the discrete space, the difference being that instead of FP we have the Master equation

For a Markov process over discrete states, we can write the master equation as

$$\frac{\partial P_j}{\partial t} = \sum_{k \neq j} \left[\Gamma_{k \to j} P_k - \Gamma_{j \to k} P_j \right]. \tag{3.41}$$

3.4 Quantum Mechanics, Control and loss functions

3.4.1 Continuous space

3.4.2 Discrete space

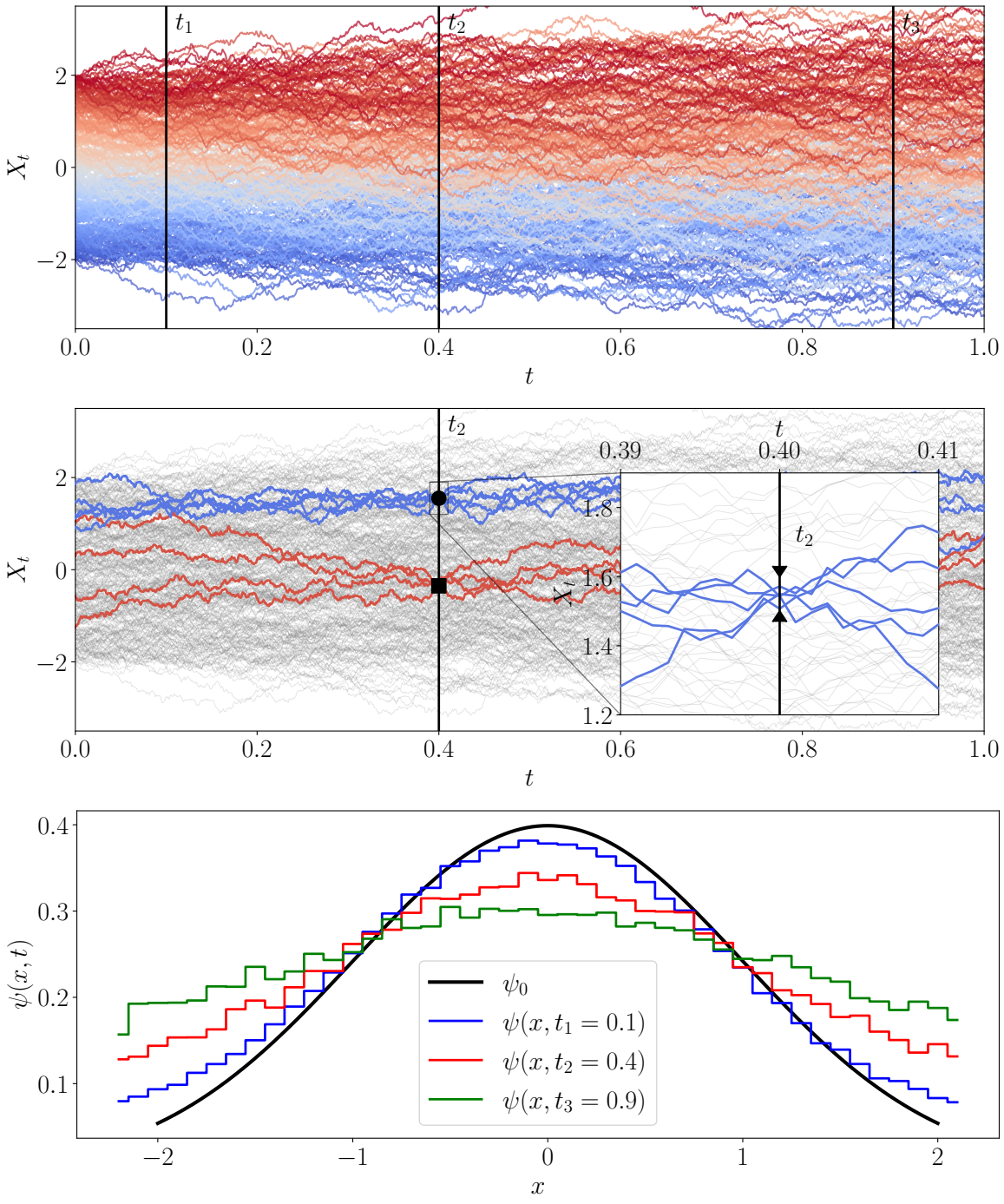


Fig. 3.4 **Feynman-Kac for a free particle in 1D. top:** $N = 400$ Brownian walks starting from different x_0, the color signifies initial position. In order to evaluate ψ between $x - \frac{\delta x}{2}$ and $x + \frac{\delta x}{2}$ at some time t we must first find Brownian paths that end there. **middle:** The paths that pass through at $x \in (1.5, 1.6)$ (blue) and through $x \in (-0.4, -0.3)$ (red) are colored, others are left in grey. **bottom:** Time evolution of the initial condition $\psi_0 = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$, by estimating $\mathbb{E}[\psi(X_0, 0)]$ from the filtered paths at each timestep.

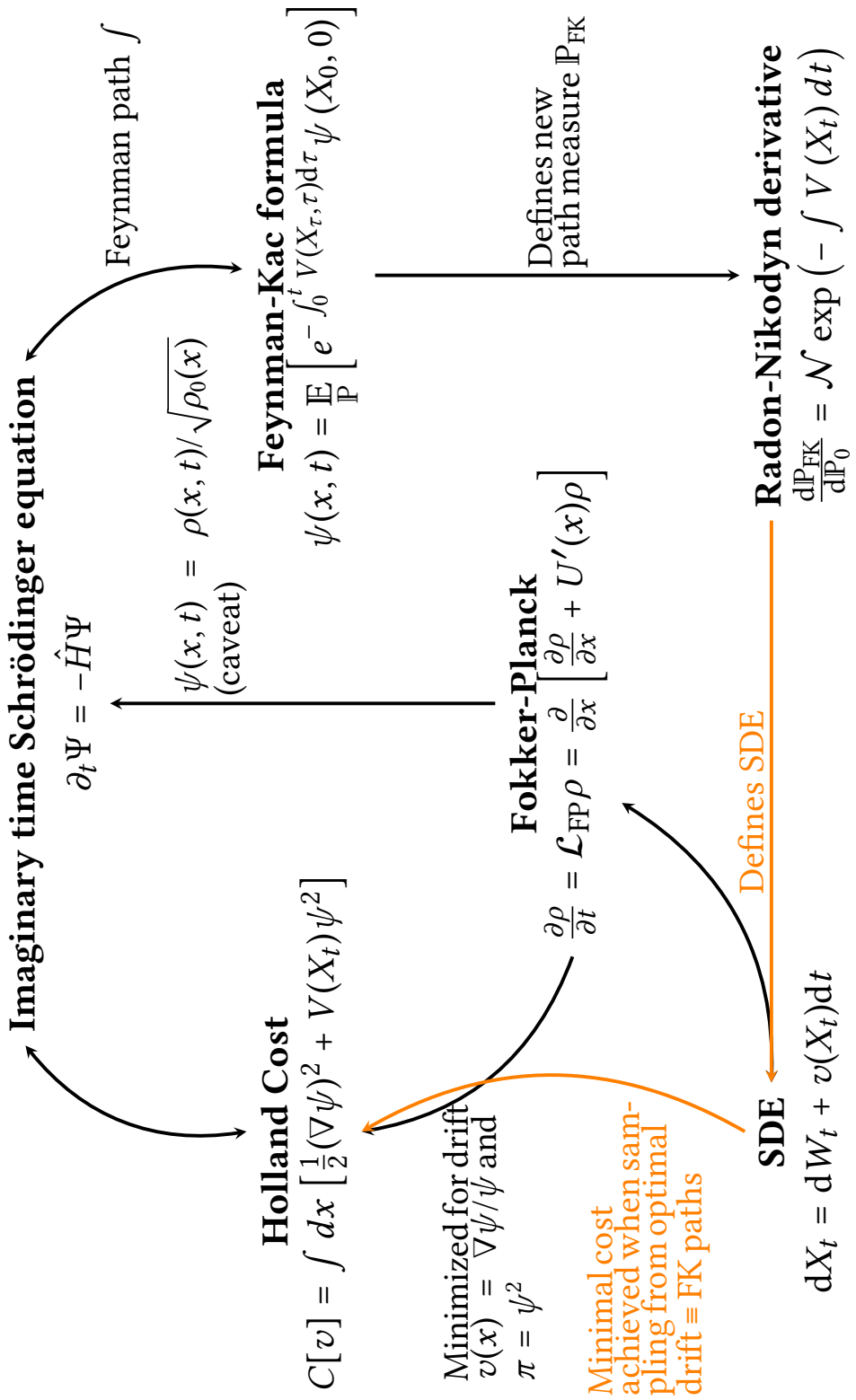


Fig. 3.6 **QM, stochastic processes and optimal control**

Chapter 4

Machine Learning

4.1 Overview of ML approaches to the Quantum many-body problem

4.1.1 Neural Network Ansatzes

-
-
-
-

4.2 Neural Networks

4.2.1 Convolutional Neural Networks

Chapter 5

Methodology

5.1 Monte Carlo Importance Sampling

The most common application of Monte Carlo methods is evaluation of integrals in high dimensional space. There MC has a distinct advantage over quadrature methods, as the statistical error decreases with the square root of samples irregardless of the dimensionality of the problem. Integrals of a function $g(\mathbf{R})$

$$I = \int g(\mathbf{R})\mathrm{d}\mathbf{R}, \tag{5.1}$$

where $\mathbf{R}$ is the *configuration* of the system or simply a *walker*, can be integrated by use of an *importance function* $\mathrm{P}(\mathbf{R})$, where $\int d\mathbf{R}\mathrm{P}(\mathbf{R}) = 1$ and $\mathrm{P}(\mathbf{R}) \geq 0$. The integral can be rewritten in the form

$$\int g(\mathbf{R})\mathrm{d}\mathbf{R} = \int \frac{g(\mathbf{R})}{\mathrm{P}(\mathbf{R})}\mathrm{P}(\mathbf{R})\mathrm{d}\mathbf{R} = \int f(\mathbf{R})\mathrm{P}(\mathbf{R})\mathrm{d}\mathbf{R}, \tag{5.2}$$

where $f(\mathbf{R}) = g(\mathbf{R})/\mathrm{P}(\mathbf{R})$. The importance function $\mathrm{P}(\mathbf{R})$ can be interpreted as a probability density. If we generate an infinite number of random uncorrelated configurations $\mathbf{R}_m$ from the distribution $\mathrm{P}(\mathbf{R})$, the sample average is a good estimator of the integral I

$$I = \lim_{M\to\infty}\left\{\frac{1}{M}\sum_{m=1}^{M} f(\mathbf{R}_m)\right\}, \tag{5.3}$$

and for an approximation with a finite number of samples

$$I \approx \frac{1}{M}\sum_{m=1}^{M} f(\mathbf{R}_m). \tag{5.4}$$

Under conditions where the central limit theorem holds [11], the estimator is normally distributed with variance σ_f^2/M, which can also be estimated from the samples as

$$\frac{\sigma_f^2}{M} \approx \frac{1}{M(M-1)} \sum_{m=1}^{M} \left[f(\mathbf{R}_m) - \frac{1}{M} \sum_{n=1}^{M} f(\mathbf{R}_n) \right]^2 . \tag{5.5}$$

5.2 Metropolis-Hastings Algorithm

The integration technique from the previous section relies on our ability to obtain samples from a probability distribution $\mathrm{P}(\mathbf{R})$. In the case of QMC these distributions are high-dimensional and cannot be directly sampled from. Moreover their normalisations are usually not known. The Metropolis-Hastings algorithm [14], see Algorithm 1, avoids direct sampling from the distribution $\mathrm{P}(\mathbf{R})$ and is insensitive to its normalisation. It uses a Markov process whose stationary distribution $\pi(\mathbf{R})$ is the same as $\mathrm{P}(\mathbf{R})$ to generate a sequence of configurations $\{\mathbf{R}_n\}_\mathrm{P}$ that are drawn from $\mathrm{P}(\mathbf{R})$. A Markov process is completely defined with its transition probability $\mathrm{P}(\mathbf{R} \longrightarrow \mathbf{R}')$, which is the probability of transitioning from state $\mathbf{R}$ to state $\mathbf{R}'$. For the process to have a unique stationary distribution two conditions must be met, the process must be *ergodic* and it must obey *detailed balance*

$$\mathrm{P}(\mathbf{R})\mathrm{P}(\mathbf{R} \longrightarrow \mathbf{R}') = \mathrm{P}(\mathbf{R}')\mathrm{P}(\mathbf{R}' \longrightarrow \mathbf{R}), \tag{5.6}$$

rewritten as

$$\frac{\mathrm{P}(\mathbf{R})}{\mathrm{P}(\mathbf{R}')} = \frac{\mathrm{P}(\mathbf{R}' \longrightarrow \mathbf{R})}{\mathrm{P}(\mathbf{R} \longrightarrow \mathbf{R}')}. \tag{5.7}$$

The right transition probability $\mathrm{P}(\mathbf{R} \longrightarrow \mathbf{R}')$ is not known, but we can express it with a trial move transition probability $\mathrm{T}(\mathbf{R} \longrightarrow \mathbf{R}')$ which we sample and acceptance probability $\mathrm{A}(\mathbf{R} \longrightarrow \mathbf{R}')$ as

$$\mathrm{P}(\mathbf{R} \longrightarrow \mathbf{R}') = \mathrm{T}(\mathbf{R} \longrightarrow \mathbf{R}')\mathrm{A}(\mathbf{R} \longrightarrow \mathbf{R}'). \tag{5.8}$$

For equation (5.7) to hold, the acceptance probability must be

$$A\left(\mathbf{R} \longrightarrow \mathbf{R}'\right) = \min\left(1, \frac{\mathrm{T}\left(\mathbf{R}' \longrightarrow \mathbf{R}\right)\mathrm{P}\left(\mathbf{R}'\right)}{\mathrm{T}(\mathbf{R} \longrightarrow \mathbf{R}')\mathrm{P}(\mathbf{R})}\right). \tag{5.9}$$

Thus to sample from any probability distribution we need only have the ability to calculate probabilities $\mathrm{P}(\mathbf{R})$ and to sample from a trial transition probability $\mathrm{T}(\mathbf{R} \longrightarrow \mathbf{R}')$. The efficiency of the algorithm depends on the amount of trial moves that we reject. All trial moves would

be accepted if $T(\mathbf{R} \longrightarrow \mathbf{R}') = P(\mathbf{R}')$, which would just mean sampling from P directly and is the very problem we are trying to solve with Metropolis-Hastings.

Algorithm 1: Metropolis-Hastings

Result: A set of configurations $\{\mathbf{R}_n\}_P$ sampled from P
Initialize walker at random configuration $\mathbf{R}$;
while *no. samples less than N* **do**
  Generate new configuration $\mathbf{R}'$ with transition probability $T(\mathbf{R} \longrightarrow \mathbf{R}')$;
  Accept the move $(\mathbf{R} \longrightarrow \mathbf{R}')$ with probability
  $A\left(\mathbf{R} \longrightarrow \mathbf{R}'\right) = \min\left(1, \frac{T(\mathbf{R}' \longrightarrow \mathbf{R})P(\mathbf{R}')}{T(\mathbf{R} \longrightarrow \mathbf{R}')P(\mathbf{R})}\right)$;
  Append $\mathbf{R}$ to the set of configuration;
end

5.3 Gradient based optimisation

5.3.1 Gradient estimation

- **TODO:** Rewrite with a more general tone, do not only talk about the ELBO. Still use Mohammeds review! It is very good.

In order to perform gradient descent on the ELBO objective, we need to be able to evaluate its gradients with respect to parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. Taking the gradient w.r.t generative parameters $\boldsymbol{\theta}$ is straightforward, because we can change the order of the expectation operator and the gradient, leaving us with

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) &= \nabla_{\boldsymbol{\theta}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}) - \log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right] \\ &\simeq \nabla_{\boldsymbol{\theta}}\left(\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}) - \log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right) \\ &= \nabla_{\boldsymbol{\theta}}\left(\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\right), \end{aligned} \tag{5.10}$$

where $\simeq$ denotes an unbiased estimator. This reversing of the order of operations is not possible when taking gradients w.r.t variational parameters $\boldsymbol{\phi}$ because the expectation $\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}$ is performed w.r.t the approximate posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$. The gradient could be estimated with a vanilla Monte Carlo estimator, but it has very high variance and is not practical [21].

The problem of stochastic gradient estimation of an expectation of a function is a well studied problem that transcends machine learning and has a variety of applications [9, 36]. Different estimators differ in from and their properties, variance being one of the most important. In their review [25] Mohamed et al. categorise MC gradient estimators into three categories

Score-function estimator

Score-function estimator: The score function is a logarithm of a probability distribution w.r.t to distributional parameters. It can be used as a gradient estimator

$$\begin{aligned} \nabla_{\theta} \mathbb{E}_{p_{\theta}(\mathbf{x})}[f(\mathbf{x})] &= \nabla_{\theta} \int p_{\theta}(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \\ &= \mathbb{E}_{p_{\theta}(\mathbf{x})}[f(\mathbf{x}) \nabla_{\theta} \log p_{\theta}(\mathbf{x})]. \end{aligned} \tag{5.11}$$

The score-function estimator is compatible with any cost function, it requires that the measure $p_{\theta}(\mathbf{x})$ is differentiable and easy to sample. Very importantly it is applicable to both discrete and continuous distribution, but has a drawback of having high variance.

Pathwise estimator

Continuous distributions can be sampled either directly by generating samples from the distribution $p_{\theta}(\mathbf{x})$ or indirectly, by sampling from a simpler base distribution $p(\boldsymbol{\epsilon})$ and transforming the variate through a deterministic path $g_{\theta}(\boldsymbol{\epsilon})$. Using this, it is possible to move the source of randomness in such a way that the objective is differentiable. In essence this approach pushes the parameters of the measure into the cost function which is then differentiated. The estimator is

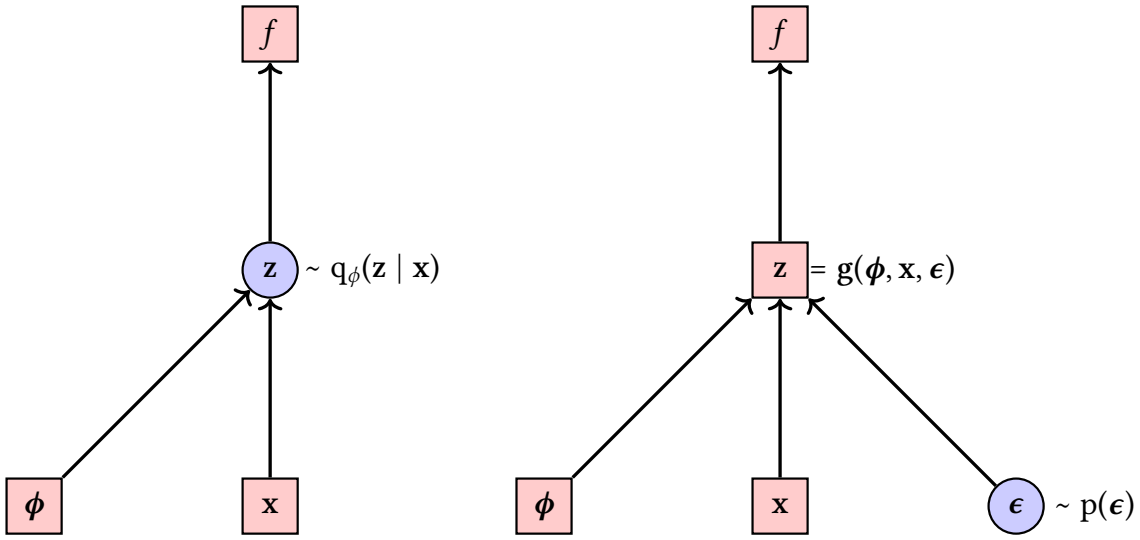
$$\begin{aligned} \nabla_{\theta} \mathbb{E}_{p_{\theta}(\mathbf{x})}[f(\mathbf{x})] &= \nabla_{\theta} \int p_{\theta}(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \\ &= \nabla_{\theta} \int p(\boldsymbol{\epsilon}) f(g_{\theta}(\boldsymbol{\epsilon})) d\boldsymbol{\epsilon} \\ &= \mathbb{E}_{p(\boldsymbol{\epsilon})}[\nabla_{\theta} f(g_{\theta}(\boldsymbol{\epsilon}))]. \end{aligned} \tag{5.12}$$

This was the gradient estimator originally used in the VAE implementation [21] there named as the *reparametrization trick*, see also Figure 5.1. In many cases the transformation paths are so simple they can be implemented in one line of code, referred to as *one-liners*. The pathwise-estimator can only be used on differentiable cost functions, but is easy to implement and crucially has lower variance than the score-function estimator.

Measure-valued gradient estimator

Which exploits the properties of signed-measures, is beyond the scope of this report.

5.3.2 Automatic differentiation



(a) Original (b) Reparametrized

Fig. 5.1 The reparametrization trick, adapted from [20]. The stochasticity of the $\mathbf{z}$ node is pushed out into a separate input to the same node, resulting in deterministic gradients w.r.t $\boldsymbol{\phi}$ through the node.

5.4 *Qptimal sampling*: optimal sampling in lattice models

Chapter 6

Results

6.1 Single Particle on a Lattice

$$\partial_t \psi_j = \frac{1}{2}\left[\psi_{j+1} + \psi_{j-1} - 2\psi_j\right] + V_j \psi_j \tag{6.1}$$

$$\psi_j(t) = \mathop{\mathbb{E}}_{X\sim\text{SRW with } X_t=j}\left[\exp\left(-\int_0^t V(X_\tau,\tau)\,\mathrm{d}\tau\right)\psi_{X_0}(0)\right] \tag{6.2}$$

6.2 Transverse-field Ising model

6.3 Heisenberg model

Heisenberg ferromagnet

$$\hat{H}_\mathrm{F} = -\frac{1}{2}\sum_j\left[\hat{\sigma}_j^x\hat{\sigma}_{j+1}^x + \hat{\sigma}_j^y\hat{\sigma}_{j+1}^y + \hat{\sigma}_j^z\hat{\sigma}_{j+1}^z\right] \tag{6.3}$$

The XY model.

$$\begin{aligned}\hat{H}_{XY} = -\sum_j\left[\hat{\sigma}_j^x\hat{\sigma}_{j+1}^x + \hat{\sigma}_j^y\hat{\sigma}_{j+1}^y\right] &= H_\mathrm{F} + \frac{1}{2}\sum_j\hat{\sigma}_j^z\hat{\sigma}_{j+1}^z \\ &= -\mathcal{W} + \sum_j\left[n_j\left(1-n_{j+1}\right) + n_{j+1}\left(1-n_j\right)\right]\end{aligned} \tag{6.4}$$

$$\psi_{s_{1:N}}(t) = \mathop{\mathbb{E}}_{\Sigma_{[0,t]} \sim \mathrm{SEP} \text{ with } \Sigma_t = s_{1:N}} \left[\exp\left(-\int_0^t dt' \sum_j \left[n_j \left(1 - n_{j+1}\right) + n_{j+1} \left(1 - n_j\right) \right] \right) \psi_{\Sigma_0}(0) \right] \tag{6.5}$$

6.4 Bose-Hubbard model

Chapter 7

Conclusions

7.1 Direction for further work

7.2 Remarks

References

[1] Albeverio, S., Ho/egh-Krohn, R., and Streit, L. (1977). Energy forms, hamiltonians, and distorted brownian paths. *Journal of Mathematical Physics*, 18(5):907–917.

[2] Anderson, J. B. (1975). A random-walk simulation of the schrödinger equation: H+ 3. *The Journal of Chemical Physics*, 63(4):1499–1503.

[3] Austin, B. M., Zubarev, D. Y., and Lester Jr, W. A. (2012). Quantum monte carlo and related approaches. *Chemical reviews*, 112(1):263–288.

[4] Carleo, G., Becca, F., Sanchez-Palencia, L., Sorella, S., and Fabrizio, M. (2014). Light-cone effect and supersonic correlations in one-and two-dimensional bosonic superfluids. *Physical Review A*, 89(3):031602.

[5] Carleo, G., Becca, F., Schiró, M., and Fabrizio, M. (2012). Localization and glassy dynamics of many-body quantum systems. *Scientific reports*, 2(1):1–6.

[6] Ceperley, D., Chester, G. V., and Kalos, M. H. (1977). Monte carlo simulation of a many-fermion study. *Physical Review B*, 16(7):3081.

[7] Ceperley, D. M. (1995). Path integrals in the theory of condensed helium. *Reviews of Modern Physics*, 67(2):279.

[8] Ceperley, D. M. and Alder, B. J. (1980). Ground state of the electron gas by a stochastic method. *Physical review letters*, 45(7):566.

[9] Chriss, N. A. and Chriss, N. (1997). *Black Scholes and beyond: option pricing models*. McGraw-Hill.

[10] Durrett, R. (2019). *Probability: theory and examples*, volume 49. Cambridge university press.

[11] Foulkes, W., Mitas, L., Needs, R., and Rajagopal, G. (2001). Quantum monte carlo simulations of solids. *Reviews of Modern Physics*, 73(1):33.

[12] Gubernatis, J., Kawashima, N., and Werner, P. (2016). *Quantum Monte Carlo Methods: Algorithms for Lattice Models*. Cambridge University Press.

[13] Halmos, P. R. (2013). *Measure theory*, volume 18. Springer.

[14] Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*.

[15] Hohenberg, P. and Kohn, W. (1964). Inhomogeneous electron gas. *Physical review*, 136(3B):B864.

[16] Ido, K., Ohgoe, T., and Imada, M. (2015). Time-dependent many-variable variational monte carlo method for nonequilibrium strongly correlated electron systems. *Physical Review B*, 92(24):245106.

[17] Kac, M. (1949). On distributions of certain wiener functionals. *Transactions of the American Mathematical Society*, 65(1):1–13.

[18] Kalos, M. (1962). Monte carlo calculations of the ground state of three-and four-body nuclei. *Physical Review*, 128(4):1791.

[19] Kalos, M. (1966). Stochastic wave function for atomic helium. *Journal of Computational Physics*, 1(2):257–276.

[20] Kingma, D. P. (2017). Variational inference & deep learning: A new synthesis.

[21] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[22] Kohn, W. (1999). Nobel lecture: Electronic structure of matter—wave functions and density functionals. *Reviews of Modern Physics*, 71(5):1253.

[23] Léonard, C. (2014). Some properties of path measures. In *Séminaire de Probabilités XLVI*, pages 207–230. Springer.

[24] McMillan, W. L. (1965). Ground state of liquid he 4. *Physical Review*, 138(2A):A442.

[25] Mohamed, S., Rosca, M., Figurnov, M., and Mnih, A. (2020). Monte carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62.

[26] Needs, R., Towler, M., Drummond, N., Lopez Rios, P., and Trail, J. (2020). Variational and diffusion quantum monte carlo calculations with the casino code. *The Journal of chemical physics*, 152(15):154106.

[27] Nelson, E. (1967). Dynamical theories of brownian motion, princeton univ. *Press, Princeton, NJ*.

[28] Norris, J. R. and Norris, J. R. (1998). *Markov chains*. Number 2. Cambridge university press.

[29] Prokof'Ev, N., Svistunov, B., and Tupitsyn, I. (1998). Exact, complete, and universal continuous-time worldline monte carlo approach to the statistics of discrete quantum systems. *Journal of Experimental and Theoretical Physics*, 87(2):310–321.

[30] Rasmussen, C., Williams, C., Press, M., Bach, F., and (Firm), P. (2006). *Gaussian Processes for Machine Learning*. Adaptive computation and machine learning. MIT Press.

[31] Reynolds, P. J., Tobochnik, J., and Gould, H. (1990). Diffusion quantum monte carlo. *Computers in Physics*, 4(6):662–668.

[32] Rogers, L. C. G. and Williams, D. (1994). Diffusions, markov processes and martingales, volume 1: Foundations. *John Wiley & Sons, Ltd., Chichester*, 7.

[33] Rogers, L. C. G. and Williams, D. (2000). *Diffusions, Markov processes and martingales: Volume 2, Itô calculus*, volume 2. Cambridge university press.

[34] Salamon, D. (2016). *Measure and integration*. European Mathematical Society.

[35] Särkkä, S. and Solin, A. (2019). *Applied stochastic differential equations*, volume 10. Cambridge University Press.

[36] Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al. (2020). Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609.

[37] Skylaris, C.-K., Haynes, P. D., Mostofi, A. A., and Payne, M. C. (2005). Introducing onetep: Linear-scaling density functional simulations on parallel computers. *The Journal of chemical physics*, 122(8):084119.

[38] Vosko, S. H., Wilk, L., and Nusair, M. (1980). Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Canadian Journal of physics*, 58(8):1200–1211.

Appendix A

Hyperparameters

Appendix B

Fixed-Node Feynman-Kac formula

Appendix C

Additional results

Appendix D

Additional Derivations