THE UNIVERSITY
of EDINBURGH

Accelerated Sampling Schemes for High Dimensional Systems

Anton Martinsson

Doctor of Philosophy
University of Edinburgh
February 2020

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Anton Martinsson*)

4/4/2020

Till Marie, Niklas och Marianna.

Lay Summary

We construct and analyse accelerated sampling schemes for high dimensional systems. The development of these methods is fundamental for effective computational study of a large class of problems in statistics, statistical physics, chemistry and engineering. Models in these different areas associate a probabilistic likelihood to the values of the variables of the system. For example, a protein molecule may fold into one of several natural shapes which has a lower energy than others. The aim of accelerated schemes is to enhance the exploration of the most likely states that will be found under certain conditions. We focus on the design of methods based on the discretization of stochastic differential equations which result in sequential iterative procedures that generate state sequences when implemented on a computer.

Of particular interest are methods that use temperature as a way of enhancing exploration of likely states. A contribution of this thesis is the design of numerical methods based on simulated tempering in the "infinite switch limit", in which the temperature is treated as a dynamical variable that fluctuates rapidly during the simulation. This approach can be generalised to allow different characteristics to be varied during simulation; for example we develop a novel approach to constant pressure sampling and couple it with this idea. We also extend infinite switch dynamics to a general case and consider infinite switch sampling also of the pressure.

Abstract

In this thesis we discuss accelerated sampling schemes for high dimensional systems, for example molecular dynamics (MD). The development of these methods is fundamental to the effective study of a large class of problems, for which traditional methods converge slowly to the system's underlying invariant probability distribution. Due to the complexity of the landscape defined by an energy function (or, in statistical models, the log likelihood of the target probability density), the exploration of the probability distribution is severely restricted. This can have detrimental effects on the conclusions drawn from numerical experiments when potentially important states and solutions are absent in the examination of the results as a consequence of poor sampling.

The aim of accelerated sampling schemes is to enhance the exploration of the invariant measure by improving the rate of convergence to it. In this work, we first focus our attention on numerical methods based on canonical sampling by studying Langevin dynamics, for which the convergence is accelerated by extending the phase-space. We introduce a scheme based on simulated tempering which makes temperature into a dynamical variable and allows switching the temperature up or down during the exploration in such a way that the target probability distribution can be easily obtained from the extended distribution. We show that this scheme is optimal when operated in the infinite switch limit. We discuss the limitations of this method and demonstrate the excellent exploratory properties of it for a moderately complicated biomolecule, alanine-12.

Next, we derive a novel approach to constant pressure simulation that forms the basis for a family of pure Langevin barostats. We demonstrate the excellent numerical performance of Lie-Trotter splitting schemes for these systems and the superior accuracy and precision of the simultaneous temperature and pressure control in comparison to currently available schemes. The scientific importance of this method lies in the ability to control the simulation and to make better predictions for applications in both materials modelling and drug design. We demonstrate this method in simulations of state transitions in crystalline materials using the "Mercedes Benz" potential.

In a final contribution, we extend the infinite switch schemes to incorporate a general class of collective variables. In particular this allows for tempering in both temperature and pressure when combined with our new barostat. We conclude this thesis by presenting a numerical study of the computational prospects of these methods.

Acknowledgements

It is difficult to sum up the large number of people that influences a mathematical and scientific education that culminates in the decision to study for a PhD in mathematics. First, I must thank my undergraduate supervisors from Manchester, Helen Gleeson and Matthias Heil, that shaped my initial experience of academic life and were elemental in encouraging me to continue my studies in Edinburgh. I foremost thank my advisor Ben Leimkuhler whose perspective on mathematics and numerical analysis molded my own understanding of where an applied mathematician's role fits within the larger scientific community. I am grateful for the large number of people he has introduced me to and the fruitful collaborations these generated. From these collaborations I must first thank Eric Vanden-Eijnden for his influence on much of my work and whose tenacity helped push my mathematical understanding further. From my time at Duke I am grateful to Jianfeng Lu and Jonathan Mattingly who showed me the power of non-intimidating mathematical presentation.

I want to thank Gabriel Stoltz and Kostas Zygalakis for their helpful discussions and Michela Ottobre who has forged my understanding in stochastic analysis throughout my studies. I thank Charles Matthews for his Bayesian sampling example detailed in Section 1.3.1 and Greg Herschlag for his knowledge of C++. I would also like to thank Zofia Trstanova and Matthias Sachs whom without I never would have made it.

Contents

Chapter 1

Introduction and Fundamentals

This thesis focuses on the study of numerical sampling methods for dynamical systems with a large number of degrees of freedom. Molecular dynamics (MD) is an example of one such problem, which finds widespread use in much of modern computational science and is the main application on which we test the methods developed herein. Additional applications where these techniques are relevant include large dimensional statistical models such as deep neural networks and Bayesian inference.

The development of accelerated sampling schemes, particularly in the MD setting, has matured over many years and the techniques we study in this thesis are based on the incorporation of temperature, for which the physical interpretation is clearly understood through thermodynamics and statistical mechanics. Drawing on the wide body of work on molecular dynamics, which also represents the most immediate application for our work, we present most of our results in the molecular modelling context. This does not exclude the application of our results in the setting of statistical models. We illustrate this potential by presenting a particular application to political science, namely the identification by statistical means of the presence of gerrymandering in the design of political constituencies. In the last chapter of the thesis we explore a general framework of accelerated sampling schemes which would have potential application in a general statistical setting.

1.1 Molecular Dynamics

The use of computer simulations in statistical physics and molecular dynamics (MD) is refereed to as *"in silico"* experimentation and they are used in a large number of fields [1, 2, 3, 4]. The numerical methods used to study both the microscopic and macroscopic properties of chemical compounds and materials have become an integral part of modern research, where they help to guide the development and utilization of more expensive *"wet lab"* experiments [5].

The practical implementation of an MD experiment is often very challenging as the number of degrees of freedom may range into the billions. These types of simulations require the use of large high performance computing (HPC) clusters, recently often also employing hundreds of GPUs (graphical processing unit). The term "*curse of dimensionality*" [6] has been coined to refer to the superpolynomial increase in computational complexity in relation to system size and such problems must be jointly solved by both the development of faster CPU/GPUs and better numerical methods. Because of this simple fact, the limitation in the development of a new method becomes its associated computational cost which, by default, eliminates a large class of numerical methods e.g

quadrature. For very high dimensional sampling problems the only realistic alternative is to use sampling schemes such as discretized Langevin dynamics, but such methods are also subject to substantial computational difficulty.

A widely studied problem in MD is the folding of proteins in water, which is a phenomena that takes place typically on a timescale measure in milliseconds or longer. Because of the need to resolve atomic vibrational motions, the study of such processes would require around 10^{12} steps or more. This is at the limit of current computer capabilities and computations do sometimes require months of simulation time, with slower processes (often the timescale is not fully understood) simply beyond our reach. The computational challenge in molecular dynamics is significant and the methods used must be accurate, stable and robust for billions of timesteps.

1.1.1 A Microscopic Perspective on Macroscopic Properties

Above we have discussed some of the challenges associated with the study of molecular systems. In general we describe such systems in terms of N point particles (atoms) each moving in d spatial dimensions, characterized by their position $\boldsymbol{q} = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N)$ and momentum $\boldsymbol{p} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N)$ with a total number of degrees of freedom $n = d \cdot N$. Denote the domain of $\boldsymbol{q}$ as Ω, where $\boldsymbol{q} \in \Omega \subseteq \mathbb{R}^n$ and $\boldsymbol{p} \in \mathbb{R}^n$. The state of a MD system is represent as a point in *phase space* $(\boldsymbol{q}, \boldsymbol{p}) \in \Omega \times \mathbb{R}^n$. This state is known as the *microscopic description* and the knowledge of an ensemble of such states is called the *macroscopic description* of the system. Statistical physics is concerned with the study of macroscopic properties of these systems, i.e. properties that depended on, in some form, the macroscopic description, and molecular dynamics is simply a tool to generate a collection of these microscopic states.

It is well established that systems, on a macroscopic level, behave according to a definite set of rules known as the *laws of thermodynamics*. These laws are a series of statements for a system at *thermal equilibrium*. Temperature is related to the concept of hot and cold and its most important feature is that of thermal equilibrium. A thermal equilibrium can be understood in terms of the flow of energy between a hot and cold system in thermal contact; the hot system will cool whilst the cool system will warm and when the flow of heat ceases, the two systems are said to be in thermal equilibrium.

A macroscopic property of a system at equilibrium is an average, of some microscopic quantity $\varphi(\boldsymbol{q}, \boldsymbol{p})$, summed over a collection of microscopic states following some probability measure on the configuration space, $\mu(\mathrm{d}\boldsymbol{q}, \mathrm{d}\boldsymbol{p})$. This measure completely describes the macroscopic state of the system and is known as the *thermodynamic ensemble*. Formally, we express a macroscopic *observable* as the integral over this thermodynamic ensemble,

$$\mathbb{E}_\mu[\varphi] = \int_{\Omega \times \mathbb{R}^n} \varphi(\boldsymbol{q}, \boldsymbol{p}) \, \mu(\mathrm{d}\boldsymbol{q}, \mathrm{d}\boldsymbol{p}). \tag{1.1}$$

We summarize MD as a tool used to generate microscopic states, whose thermodynamic ensemble is applied to calculate some physically meaningful observable. In this thesis we are predominantly concerned with the calculation of integrals of the form (1.1), often also with $\varphi(\boldsymbol{q}, \boldsymbol{p}) = \varphi(\boldsymbol{q})$.

We finish this introduction with a discussion of the choice of domain, Ω. This should be an informed choice driven by the physically conditions that the simulation is required to mimic. To simulate the bulk of a system, a periodic domain is often introduced. This is a bounded simulation cell with side length $L > 0$ and in this case

we consider $\boldsymbol{q} \in \Omega = L\mathbb{T}^n$, where $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ denotes the unit torus. Apart from this, the most important decision is choosing the right ensemble.

1.2 Thermodynamic Ensembles

The ensemble of a system of N particles in a volume V, both fixed, in isolation with constant energy E, is referred to as the *microcanonical ensemble*. In relation to the constant quantities, the ensemble is often denoted as NVE, which we use as a shorthand notation. Likewise, other ensembles exist with their own individual physical interpretations (we always restrict our studies to systems with a constant number of particles). In a more physicaly relevant setup, we attribute the *canonical ensemble* to a system with fixed volume V but now instead at thermal equilibrium with a heat bath of infinite heat capacity, at temperature T. The shorthand notation of this system is NVT. Finally, we introduce the *isothermal-isobaric ensemble*, which mimics real world conditions closely, as it defines an ensemble of a system at thermal equilibrium with constant pressure, denoted as NPT.

The NVE ensemble relates directly to Hamiltonian dynamics and has therefore, historically, received a lot of interest. Constant energy simulations are still widely used by practitioners in MD, often merely as a consequence of the extensive use of legacy software. In recent years, constant temperature simulations have been welcomed by the MD community as these imitate conditions that are closer in nature to real life experiments. As a combined result of the simple implementation and gain in physical relevance, methods based on the application of an underlying NVT ensemble form the basis of most accelerated sampling schemes, summarised in Chapter 3.

Implementations of methods building on NPT ensembles are complicated. However, since the ensemble is the most realistic, it is also the one of greatest use to practitioners of MD. There are a vast number of methods implementing, or claiming to implement, isothermal-isobaric ensembles, some of which are of ambiguous performance and in Chapter 5 we address these issues and introduce a simple and clean implementation of an accurate sampling scheme for the NPT ensemble.

The differences between these ensembles are understood through thermodynamics as a set of statements for the equilibrium of reversible heat exchanges between systems. All the systems which are studied in this thesis are in equilibrium and both NVT and NPT describe a different type of equilibrium of greatest interest to us. These are distinguished through the use of thermodynamic potentials, which are Legendre transforms of the free energy of the system. In essence, each of the equilibrium conditions above are defined using a different set of natural variables at which a certain thermodynamic potential attains its minimum. To make these statements more specific, consider the first law of thermodynamics,

$$\mathrm{d}E = T\,\mathrm{d}S - \mathcal{P}\,\mathrm{d}V. \tag{1.2}$$

This equation describes the change in internal energy E resulting from a reversible heat exchange for a closed system with a fixed number of particles and we define the variables as follows: T is the temperature, S is the entropy, $\mathcal{P}$ is the target pressure and V is the volume. Under the condition that volume V is fixed, the definition of temperature arises as,

$$\frac{1}{T} = \left[\frac{\partial S}{\partial E}\right]_V, \tag{1.3}$$

where we use $[\,]_C$ to indicate invariant quantities. Similarly, the definition of the target

pressure, $\mathcal{P}$, derives in the same way. The thermodynamic potential that is used to represent the thermal equilibrium in the case of constant temperature is the free energy,

$$F := E - TS. \tag{1.4}$$

In this case the volume V is fixed and we say that the natural variables in which to describe such systems are: T and V. Similarly, for a system at constant pressure and temperature the natural variables are $\mathcal{P}$ and T, in which case the *Gibbs free energy* is the associated thermodynamic potential,

$$G := E + \mathcal{P}V - TS. \tag{1.5}$$

This is the Legendre transform of the *enthalpy*, $H := E + \mathcal{P}V$, and plays the same role as the free energy F i.e. for a system equilibrium specified by $(\mathcal{P}, T)$, G is minimised.

Phase transitions are a phenomena at which materials are transformed from one phase to another e.g liquid-solid and liquid-gas. These conditions are associated with either heat absorption or emission by the material. As a consequence, a phase transition is best described via the Gibbs free energy (1.5) and the *Clausius-Clapeyron equation*,

$$\frac{\mathrm{d}\mathcal{P}}{\mathrm{d}T} = -\frac{\Delta\left(\left[\frac{\partial G}{\partial T}\right]_{\mathcal{P}}\right)}{\Delta\left(\left[\frac{\partial G}{\partial \mathcal{P}}\right]_{T}\right)} = \frac{\Delta S}{\Delta V}. \tag{1.6}$$

A phase transition is detected in terms of discontinuities of the derivatives in this equation which implies that there are differences in the entropy and density between the two phases. A *first order phase transition* involves latent heat and means that the entropy changes. This implies that $[\partial G/\partial T]_{\mathcal{P}}$ is discontinuous at the phase transition conditions. We use a more detailed argument in Chapter 4 that can be interpreted as a first order phase transition.

1.2.1 NVE, Constant Energy

Let E be the total energy of a system of N particles. The microcanonical ensemble is described by a generalized (Dirac delta function) density through the relation,

$$\mu_E(\mathrm{d}\boldsymbol{q}, \mathrm{d}\boldsymbol{p}) = Z_E^{-1}\delta\left(H\left(\boldsymbol{q}, \boldsymbol{p}\right) - E\right)\mathrm{d}\boldsymbol{q}\,\mathrm{d}\boldsymbol{p}, \tag{1.7}$$

where Z_E is a normalization constant such that $\int \mu_E = 1$,

$$Z_E = \int_{\Omega\times\mathbb{R}^n} \delta\left(H\left(\boldsymbol{q}, \boldsymbol{p}\right) - E\right)\mathrm{d}\boldsymbol{q}\,\mathrm{d}\boldsymbol{p}\ . \tag{1.8}$$

This measure is understood in the sense of generalized functions through averages defined by [7],

$$\begin{aligned}\int_{\Omega\times\mathbb{R}^n} \varphi\left(\boldsymbol{q}, \boldsymbol{p}\right)\delta\left(H\left(\boldsymbol{q}, \boldsymbol{p}\right) - E\right)\mathrm{d}\boldsymbol{q}\,\mathrm{d}\boldsymbol{p}& \\ := \lim_{\Delta E\to 0}\frac{1}{2\Delta E}\int_{\Omega\times\mathbb{R}^n} \varphi\left(\boldsymbol{q}, \boldsymbol{p}\right)\mathbb{1}_{M_{E,\Delta E}}\left(\boldsymbol{q}, \boldsymbol{p}\right)\mathrm{d}\boldsymbol{q}\,\mathrm{d}\boldsymbol{p},&\end{aligned} \tag{1.9}$$

where,

$$M_{E,\Delta E} = \left\{(\boldsymbol{q}, \boldsymbol{p}) \in \Omega\times\mathbb{R}^n \mid E - \Delta E \le H\left(\boldsymbol{q}, \boldsymbol{p}\right) \le E + \Delta E\right\}. \tag{1.10}$$

The act of sampling measures of the form (1.7) i.e. proposing a collection of points $(\boldsymbol{q}, \boldsymbol{p})$ that conserves this quantity can be achieved via Hamiltonian dynamics. In the section below, we recall a few important concepts and set some of the notation for use in following Chapters.

Hamiltonian Dynamics

Assume that each particle $i \in \{1, \ldots, N\}$ has an associated mass $\boldsymbol{m}_i \in (\mathbb{R}_+ \setminus \{0\})^d$. We denote the mass matrix of this system as the diagonal matrix,

$$M = \operatorname{diag}(\boldsymbol{m}_1, \ldots, \boldsymbol{m}_N) \otimes I_d, \tag{1.11}$$

where M is an $n \times n$ diagonal matrix and I_d denotes the identity matrix in $\mathbb{R}^d$. The point particles in a MD system are subject to a force, which is the negative gradient of some (differentiable) potential energy function $U : \Omega \to \mathbb{R}$. The study and formulation of these force fields is a very active research area and are often empirical approximations e.g. [8, 9] or recently also via a data driven approach as in [10, 11]. Here, we will treat the potential energy function as given. The total microscopic energy of this system is given by the Hamiltonian,

$$H(\boldsymbol{q}, \boldsymbol{p}) = \frac{1}{2}\boldsymbol{p}^T M^{-1} \boldsymbol{p} + U(\boldsymbol{q}), \tag{1.12}$$

where $\boldsymbol{p}^T$ denotes the transpose of the vector $\boldsymbol{p}$. The particle motion is assumed to obey classical mechanics and Newton's equation of motion, which can be written in terms of the Hamiltonian (1.12) as,

$$\begin{aligned} \dot{\boldsymbol{q}} &= \nabla_{\boldsymbol{p}} H(\boldsymbol{q}, \boldsymbol{p}), \\ \dot{\boldsymbol{p}} &= -\nabla_{\boldsymbol{q}} H(\boldsymbol{q}, \boldsymbol{p}). \end{aligned} \tag{1.13}$$

Here we suppress the dependence on time t in both $\boldsymbol{q} = \boldsymbol{q}(t)$ and $\boldsymbol{p} = \boldsymbol{p}(t)$ to simplify the notation. We now recall some properties of the flow map of Hamiltonian dynamics defined as $\phi_t : \Omega \times \mathbb{R}^n \to \Omega \times \mathbb{R}^n$. From some initial condition $(\boldsymbol{q}(0), \boldsymbol{p}(0))$ we assume that ϕ_t exists for all $t \in \mathbb{R}$ and that the solution to (1.13) is $(\boldsymbol{q}(t), \boldsymbol{p}(t)) = \phi_t(\boldsymbol{q}(0), \boldsymbol{p}(0))$. The following hold for the flow ϕ_t.

1. **Energy Conservation:** For all $t \in \mathbb{R}$ under the flow ϕ_t we have,

$$H(\boldsymbol{q}(t), \boldsymbol{p}(t)) = H(\boldsymbol{q}(0), \boldsymbol{p}(0)). \tag{1.14}$$

2. **Volume Preservation:** We can represent the dynamics in (1.12) as,

$$\begin{pmatrix} \dot{\boldsymbol{q}} \\ \dot{\boldsymbol{p}} \end{pmatrix} = J \nabla H(\boldsymbol{q}, \boldsymbol{p}), \tag{1.15}$$

where J is the skew-symmetric matrix, known as the symplectic structure matrix [12]. Here,

$$J = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix}. \tag{1.16}$$

From (1.15), the flow must be volume conserving. For any measurable set $A \in$

$\Omega \times \mathbb{R}^n$ and all $t \in \mathbb{R}$ we have from Liouville's theorem,

$$\int_{\phi(A)} \mathrm{d}\boldsymbol{q}\,\mathrm{d}\boldsymbol{p} = \int_A \mathrm{d}\boldsymbol{q}\,\mathrm{d}\boldsymbol{p}. \tag{1.17}$$

3. **Symplecticity:** It is a fundamental property of the Hamiltonian maps $\{\phi_t\}_{t\in\mathbb{R}}$ that they have a symplectic group structure [13]. This is defined as,

$$\nabla\phi_t^T J \nabla\phi_t = J, \qquad \forall t \in \mathbb{R}, \tag{1.18}$$

with $\nabla\phi_t$ the Jacobian of ϕ_t.

We use the term *symplectic integration* to refer to a numerical timestepping method which respects properties 2-3 above (and only property 1 approximately).

1.2.2 NVT, Constant Temperaure

The statistical framework presented in this section is the framework considered in a large number of contributions in the numerical stochastic literature [14, 15, 16, 17, 18, 19, 20]. We refer to [21, chp. 3, 4] for more detail as the scope of this section is only to give a general overview, modelled on the discussion by Lelievre et al. [7, Sec. 1.2.3.2].

When N particles $(\boldsymbol{q}, \boldsymbol{p})$ in an isolated system are in contact with a heat bath (a body with a heat capacity much larger than that of the system), energy flows into or out of the system from the bath. This implies that the instantaneous energy of the system is no longer fixed, but fluctuates over time although it remains fixed in average; we assume that the system is in thermal equilibrium with the heat bath at temperature T.

At this equilibrium, the temperature of the system is a well defined quantity and the particles' positions in phase space are distributed according to the measure,

$$\mu_\beta(\mathrm{d}\boldsymbol{q}, \mathrm{d}\boldsymbol{p}) = Z_\beta^{-1} \exp\left[-\beta H\left(\boldsymbol{q}, \boldsymbol{p}\right)\right] \mathrm{d}\boldsymbol{q}\,\mathrm{d}\boldsymbol{p}, \tag{1.19}$$

where $\beta = 1/k_B T$ is the reciprocal temperature scaled by the Boltzmann constant k_B. The normalisation constant Z_β ensures that μ_β is a probability measure i.e. $\int \mu_\beta = 1$:

$$Z_\beta = \int_{\Omega\times\mathbb{R}^n} \mu_\beta(\mathrm{d}\boldsymbol{q}, \mathrm{d}\boldsymbol{p}), \tag{1.20}$$

and this quantity is commonly called the *partition function*. Formally, the derivation of (1.19) is the solution to a variational problem. Assume that the measure $\mu(\mathrm{d}\boldsymbol{q}, \mathrm{d}\boldsymbol{p})$ has a density $\rho\left(\boldsymbol{q}, \boldsymbol{p}\right) \in L^1\left(\Omega \times \mathbb{R}^n\right)$ w.r.t. to the Lebesgue measure where ρ must satisfy,

$$\rho\left(\boldsymbol{q}, \boldsymbol{p}\right) \geq 0, \quad \int_{\Omega\times\mathbb{R}^n} \rho\left(\boldsymbol{q}, \boldsymbol{p}\right) \mathrm{d}\boldsymbol{q}\mathrm{d}\boldsymbol{p} = 1, \quad \int_{\Omega\times\mathbb{R}^n} H\left(\boldsymbol{q}, \boldsymbol{p}\right) \rho\left(\boldsymbol{q}, \boldsymbol{p}\right) \mathrm{d}\boldsymbol{q}\mathrm{d}\boldsymbol{p} = E. \tag{1.21}$$

Here, E is some energy level and the first two terms ensure that ρ is a probability density.

The concept of thermodynamics and entropy was introduced in the late 19^{th} century with the rise of desire to design ever more efficient steam engines. This area of thermodynamics was studied by some of natural sciences' giants but most significantly,

Ludvig Boltzmann. In 1872 he wrote down a functional [22, p.126] similar to,

$$S(\rho) = -\int_{\Omega\times\mathbb{R}^n} \rho(\boldsymbol{q},\boldsymbol{p}) \log \rho(\boldsymbol{q},\boldsymbol{p}) \,\mathrm{d}\boldsymbol{q}\mathrm{d}\boldsymbol{p}, \tag{1.22}$$

which we denote as the *statistical entropy*. Using the inequality $x \log x \geq x-1$ for $x > 0$ together with the fact that ρ is a probability density, one deduces that (1.22) is non positive. The canonical measure (1.19) is then deduced from solving the optimization problem,

$$\sup_{\rho\in L^1(\Omega\times\mathbb{R}^n)} \left\{ S(\rho) \;\middle|\; \rho \geq 0,\ \int_{\Omega\times\mathbb{R}^n} \rho = 1,\ \int_{\Omega\times\mathbb{R}^n} H\rho = E \right\}. \tag{1.23}$$

In fact, it can be shown that the unique solution to this problem is exactly (1.19).

Langevin Dynamics

In contrast to deterministic constant energy simulations via Hamiltonian dynamics, constant temperature sampling can be conducted using a selection of different methods. The approach discussed in this section and that we also focus on in this thesis is *Langevin dynamics*. The two commonly used approaches in the derivation of the *generalised Langevin equation* (GLE) are the method due to Zwanzig [23] or by coupling a field with Hamiltonian equations (see e.g Pavliotis [18, ch.8] or Rey-Bellet [24]). For a discussion on the properties of the GLE equation we refer to [25].

We exclusively consider a memory kernel with vanishing noise correlation, which is given by the constant $\gamma \in [0,\infty)$. As a consequence we have for a general $(\boldsymbol{q},\boldsymbol{p}) \in \Omega \times \mathbb{R}^n$,

$$\begin{aligned} \mathrm{d}\boldsymbol{q} &= M^{-1}\boldsymbol{p}\,\mathrm{d}t, \\ \mathrm{d}\boldsymbol{p} &= -\nabla_{\boldsymbol{q}} U(\boldsymbol{q})\,\mathrm{d}t - \gamma M^{-1}\boldsymbol{p}\,\mathrm{d}t + \sigma\,\mathrm{d}\boldsymbol{W}, \end{aligned} \tag{1.24}$$

where $\sigma\sigma^T = 2\gamma\beta^{-1}$ and $\mathrm{d}\boldsymbol{W}$ is a standard n-dimensional Wiener process.

1.2.3 NPT, Constant Temperature and Pressure

This ensemble is the most complex, of the frameworks we consider, to describe of the main ones used in molecular modelling, but it is also the one that closely resembles real world laboratory conditions. It is known in the statistical physics literature as the *isothermal-isobaric ensemble*. Under these conditions particles are not confined to a domain of fixed volume, but instead to a dynamic simulation cell with fluctuating volume. To be precise, let $V \in (0,\infty)$ be the volume of a cubic cell with side L in $d = 3$ with target pressure $\mathcal{P}$. We define the domain of the position of the N particles as,

$$\Omega(L) = \left\{ \boldsymbol{x} \in \mathbb{R}^3 \mid L^{-1}\boldsymbol{x} \in \mathbb{T}^3 \right\}^N. \tag{1.25}$$

Note the use of $\mathbb{T}$ which implies that we consider a periodic simulation cell, mimicking bulk conditions. The cell can also be extended to be fully flexible in each direction instead of the fixed-shape cube, as in this situation.

As mentioned above, thermodynamic potentials are used to identify equilibrium positions, which for the constant temperature sampling we recall as the minimum of the free energy. In this case the relevant thermodynamic potential to use is the Gibbs

potential (1.5), which is maximised for the measure given by,

$$\mu_{\beta,\mathcal{P}}\left(\mathrm{d}\boldsymbol{q},\mathrm{d}\boldsymbol{p},\mathrm{d}L\right)=Z_{\beta,\mathcal{P}}^{-1}\exp\left[-\beta\left(H\left(\boldsymbol{q},\boldsymbol{p}\right)+\mathcal{P}L^{3}\right)\right]\mathbb{1}_{\boldsymbol{q}\in\Omega(L)}\,\mathrm{d}\boldsymbol{q}\,\mathrm{d}\boldsymbol{p}\,\mathrm{d}L, \tag{1.26}$$

where $Z_{\beta,\mathcal{P}}$ is the normalization constant,

$$Z_{\beta,\mathcal{P}}=\int_{(0,\infty)}\mathrm{d}L\int_{\Omega(L)\times\mathbb{R}^{n}}\mathrm{d}\boldsymbol{q}\mathrm{d}\boldsymbol{p}\;\exp\left[-\beta\left(H\left(\boldsymbol{q},\boldsymbol{p}\right)+\mathcal{P}L^{3}\right)\right]. \tag{1.27}$$

Barostat Dynamics

The treatment of the set of equations for the fully flexible cell is given in Chapter 5. Equation (1.26) seems deceptively easy to implement in practice and at first glance one might like to think of it as a special case of the Langevin equations. This is not entirely true and care must be taken in addressing some of the problems arising here.

Because of the interdependence between the domain of the particle positions $\Omega(L)$ and the simulation cell side length L, a trivial expression of a Langevin process is not possible. This was first remarked on by Anderson [26] in 1980, that proposed a mapping of the particle positions such that the new positions are defined on a domain which is independent of L. Later we will show that this subtlety has been poorly treated in the literature and we aim to give a thorough treatment of this somewhat complex situation.

1.3 Statistical Models

Just as molecular dynamics is a model for how matter behaves at a microscopic level, statistical inference models aim to predict some response $\boldsymbol{y}$ given data $\boldsymbol{x}$, by fitting some model parameters $\boldsymbol{\theta}$. The fitting of these model parameters to the data has often been treated as a convex problem where the unique minimum is obtained by application of e.g gradient descent algorithms.

In many situations, finding the globally optimal model parameters that best describe the response $\boldsymbol{y}$ given the data $\boldsymbol{x}$, is a non-convex optimization problem. As we shall discuss below, many of the challenges presenting themselves in these problems are similar to the challenges posed in molecular dynamics.

We stress that the aim of sampling in this situation is not necessarily to obtain the globally optimal $\boldsymbol{\theta}$ at the end of the simulation, but rather to acquire more information by sampling the log-likelihood function to instead obtain distributional information. This change of perspective might be exactly what is desired in certain situations, but care must be taken such that sampling techniques are not applied under circumstances where optimisation is required and vice versa.

In a more general setting statistical models could be interpreted as complex models of some real-world application for which the aim is to obtain statistics. In this situation the application of a sampling algorithm, that we assume samples some well defined distribution, helps to systematically produce representative choices of the model parameters. These models could be arbitrarily complex and the modelling choices of the system at hand become important to consider, as they ultimately influence how the model can be sampled and what techniques are suitable. We discuss one such example in terms of geo-political districting in Section 1.3.2.

1.3.1 Bayesian Inference

Many different statistical models exist, but in this section we illustrate how sampling can be used to explore the parameterisation of a Gaussian mixture model. An interesting paper on the role of Bayesian statistics in the philosophy of statistical models can be found in Lindley [27] and we refer the reader to this paper for an overview of the subject.

Bayesian inference has gained wider popularity with the increase in computational power, unlocking the potential of this powerful framework for incorporating uncertainties about both the data $p(\boldsymbol{x}|\boldsymbol{\theta})$ and the model parameters $p(\boldsymbol{\theta})$. This means that it is aptly suited as an application of sampling methods – that are designed to explore some probability density functions.

Consider a sequence of random variables $\{\boldsymbol{X}_i\}_{0\leq i<N}$ on some measurable space $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ distributed i.i.d with density $p(\cdot|\boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \mathbb{R}^d$. Assume also that the parameters $\boldsymbol{\theta}$, to the best of our knowledge, are distributed according to the prior $\pi(\boldsymbol{\theta})$. In this situation the analogue of the potential energy function U in Section 1.2.1 is the log of the posterior distribution

$$U(\boldsymbol{\theta}|\boldsymbol{X}) = \log \rho(\boldsymbol{\theta}|\boldsymbol{X}). \tag{1.28}$$

Here we use $\boldsymbol{X}$ to refer to the entire collection of data. The posterior distribution, $\rho(\boldsymbol{\theta}, \boldsymbol{X})$ encodes the uncertainty regarding the prediction that we can make given that we are initially uncertain about the data, and is in general an unknown function. The posterior distribution is only accessible through statistical inference via Bayes formula,

$$\rho(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{\pi(\boldsymbol{\theta})\Pi_{i=0}^{N-1} p(\boldsymbol{X}_i|\boldsymbol{\theta})}{\int_{\mathbb{R}^d} \pi(\boldsymbol{\theta})\Pi_{i=0}^{N-1} p(\boldsymbol{X}_i|\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}. \tag{1.29}$$

Here, the function $p(\boldsymbol{X}_i|\boldsymbol{\theta})$ is called the *statistical model* or also the *likelihood function* and should be chosen as an informed decision based on the belief that the practitioner has about the data. Many different models exist, one of the simplest being the *Gaussian mixture model*. In this model the distribution of the data is assumed to be well described by a set of Gaussians and the goal is to find the parameters $\boldsymbol{\theta} = (\boldsymbol{m}, \boldsymbol{\sigma})$ (means and standard deviations), that best represent the data.

To sample (1.28) we set up the Langevin process,

$$\begin{aligned} \mathrm{d}\boldsymbol{\theta} &= \boldsymbol{p}_\theta\,\mathrm{d}t, \\ \mathrm{d}\boldsymbol{p}_\theta &= -\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}|\boldsymbol{X})\,\mathrm{d}t - \boldsymbol{p}_\theta\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\boldsymbol{W}. \end{aligned} \tag{1.30}$$

Here we draw from the previous section on Langevin dynamics and we have also introduced the fictitious momentum $\boldsymbol{p}_\theta$, which is the conjugate momentum of the parameters $\boldsymbol{\theta}$. Additionally, we set all the dynamical parameters to unity, which is not essential; they can be incorporated to tune the exploration of the process.

To make this more explicit we consider the case in which the statistical model is given by,

$$p(\boldsymbol{X}_i|\boldsymbol{\theta}) = \sum_{j=0}^{d-1} \left(2\pi\sigma_j^2\right)^{-1/2} \exp\left[-\frac{|\boldsymbol{X}_i - m_j|^2}{2\sigma_j^2}\right]. \tag{1.31}$$

Using this equation in combination with (1.28) and (1.29), the potential is written as,

$$U\left(\boldsymbol{\theta}|\boldsymbol{X}\right) = \log \pi\left(\boldsymbol{\theta}\right) + \sum_{i=0}^{N-1} \log p\left(\boldsymbol{X}_i|\boldsymbol{\theta}\right). \tag{1.32}$$

The normalisation constant coming from Bayes formula (1.29) is irrelevant and has been omitted for clarity. This is because we are only interested in the gradient of U w.r.t. $\boldsymbol{\theta}$, which is independent of this constant.

To illustrate the use of sampling methods for models of this type we use the following simple example. We draw 30 random points from the one dimensional distribution,

$$\rho_{\text{exact}} = \frac{1}{3}\left(\mathcal{N}\left(\sigma_{e,0}, 1\right) + \mathcal{N}\left(\sigma_{e,1}, 1\right) + \mathcal{N}\left(\sigma_{e,2}, 1\right)\right), \tag{1.33}$$

such that $\{X_i\}_{0\leq i<30} \sim \rho_{\text{exact}}$ where we choose $\sigma_{e,0} = 11$, $\sigma_{e,1} = 13.5$ and $\sigma_{e,2} = 16$. For simplicity, we assume that all variances are unity, such that the only fitting parameters are three means $\boldsymbol{\theta} = (m_0, m_1, m_2)$, with the prior chosen as,

$$\pi(\boldsymbol{\theta}) = \exp\left[-\frac{(m_p - m_0)^2}{2\sigma_p}\right] + \exp\left[-\frac{(m_p - m_1)^2}{2\sigma_p}\right] + \exp\left[-\frac{(m_p - m_2)^2}{2\sigma_p}\right]. \tag{1.34}$$

Here we pick the mean as the empirical average $m_p = 13.5$ and a large standard deviation $\sigma_p = 4$. By differentiating (1.32) w.r.t. to $\boldsymbol{\theta}$, assuming that $\sigma_j = 1$ for $j = 0, 1, 2$, the dynamics given in (1.30) is used to sample 5×10^4 points from the posterior distribution $\rho(\boldsymbol{\theta}|\boldsymbol{X})$. In the left panel of Figure 1.1 we show the sorted distribution of the means which are obtained from the sampling. Note that no constraints were imposed on the order of the fitting parameters and as a consequence we sample all order permutations.

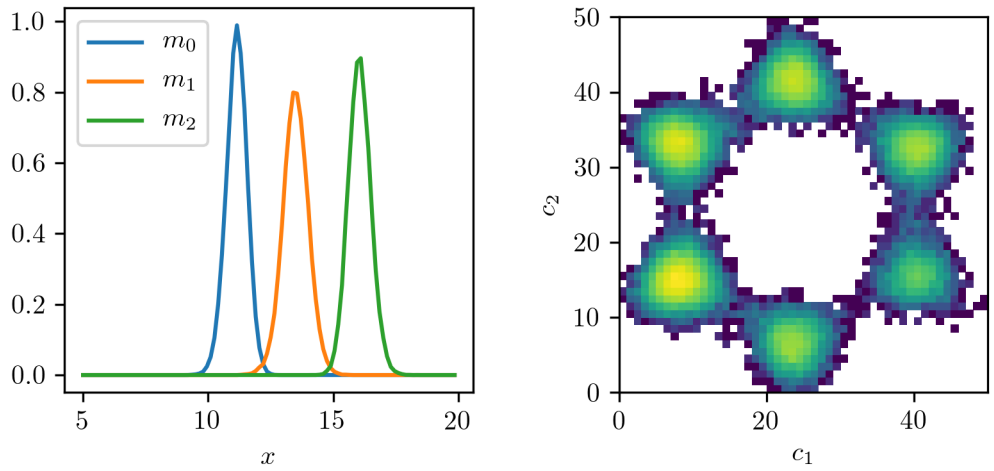

Figure 1.1: (left) Figure showing the sorted distribution of the means and their Gaussian distribution. (right) A heat map illustrating the symmetry in the distribution of the means. Here each metastable state represents a permutation in the order of the three means m_0, m_1, m_2, in the variables c_1, c_2 defined in the text.

In fact an illustrative representation in coordinates that makes these permutations

explicit, exists in the form of:

$$c_1 = m_0 - m_1, \qquad c_2 = 2m_2 - m_1 - m_0. \tag{1.35}$$

A heatmap in this representation is plotted in the right panel of Figure 1.1. This clearly shows the expected 3! possible configuration of the means and does show that each permutation is metastable, regions of high probability (yellow) are separated by regions of low probability (blue). This will be discussed in more detail in the following section as this presents one of the main challenges to sampling.

Although the example presented in this section is a simple toy problem, everything which is discussed with regards to accelerated sampling in the coming chapters, can be applied to statistical problems of this type. We again stress that we have chosen to present all the methods in the light of MD because of the clear physical interpretation of state variables and parameters such as temperature and friction etc. It can be argued that this lack of physical connection is the main disadvantage in using these types of sampling methods for statistical models, as its not clear even what order of magnitude the temperature, should have.

1.3.2 Sampling Political Districts

Gerrymandering is a complex tool often used by political parties to establish a political advantage by redrawing congressional district lines. This has huge consequences for the electorate as votes are effectively wasted. Over prolonged periods of time this could deter belief in the democratic process, leading to polarization as people feel they are not being listened to. In the US it is not, currently, illegal to gerrymander based on partisan belonging. A political advantage is in this case obtained by using historical voting data and spreading votes from ones own party over as many districts as possible, such that a large number of districts are won by a narrow margin. Simultaneously, the opponents votes are packed into a small number of districts that are lost by a very large margin. These techniques are called "cracking" and "packing", and when utilised successfully the opponent wastes a large number of votes winning few districts, whilst the own party narrowly wins many.

The processes involved in gerrymandering are complex and even detecting it, i.e. claiming that a particular election result has been gerrymandered, is problematic as the political beliefs of the person making the claim cannot be excluded. In addition, geographical differences in the physical landscape could mean that certain maps should naturally be favoured so as to spread the political influence fairly. To combat these problems, Herschlag et. al [3] quantify the fairness of the political landscape in North Carolina (NC) by sampling the space of congressional district maps.

The goal of this approach is to obtain a wide range of district maps by sampling, where all maps are required to satisfy conditions on population and shape that are loosely imposed on the system through the definition of some energy. After collecting a large number of maps one evaluates the likelihood of proposed or used maps when compared to the large ensemble of maps. To understand this approach in more detail we introduce a short list of terms,

- **Precinct:** The lowest level of electoral district that a voter interacts with directly, also the most local form of government.
- **Congressional District:** (*district*) A first past the post electoral constituency that elects a single member to congress.

- **Map:** A function mapping precincts into districts.

In North Carolina there are 13 districts and we define a map as a configuration of a 13-spin particle system, where each particle represents one of the 2690 precincts. Random configurations are not allowed and a few rules are introduced by Herschlag et. al that are used to quantify the energy of the model.

We only use a single energy term. Define the iso-parametric score as in [3], for which the ideal value is the iso-parametric score of the circle I_{ideal}. The energy term for this setup is defined as,

$$U(\boldsymbol{I}) = \sum_{j=0}^{12} \left(\frac{I_j}{I_{\text{ideal}}} - 1 \right)^2 . \tag{1.36}$$

Here $\boldsymbol{I}$ is the vector containing the iso-parametric score of all districts and I_j is the iso-parametric score of the j^{th} district.

Let the ideal population, P_{ideal}, be the total population of NC divided by 13. We call a map "compliant" when all of the 13 districts' populations deviate from this number by a maximum of 5%. In the sampling algorithm we implement this as a hard constraint and accept maps with population deviations within 25% of this value. This implies that the space of compliant maps is a subset of the space of accepted maps.

Note that this setup is not unique and differs from the one used by Herchlag et. al., otherwise our approach is identical: using the same Metropolis-Hastings algorithm, where a conflicting edge (an edge connecting two precincts with different labels) is picked and each of the two precincts connected by this edge are flipped with probability 0.5. This means that the algorithm samples the distribution with density given by,

$$\rho_\beta \propto e^{-\beta U(\boldsymbol{I})}. \tag{1.37}$$

Here, β is a fictitious parameter that imitates the role of temperature.

Below, in Figure 1.2 we illustrate the centre of mass distribution (where mass is taken as population) of three representative districts when sampled at $\beta = 5$. The heatmaps are plotted on top of the centroid locations of the precincts, where increased concentration indicate population centres. A lighter colour encodes a value of higher probability and these are interlaced with regions of low probability, represented by blue. This type of phenomena is the same as seen in Figure 1.1, only more complicated.

Sampling this type of problem is complex and requires significant knowledge and tuning of the parameters involved. Additionally, the choice of energy and temperature greatly influences the maps that are obtained and what methods that sample the resulting system well. Regardless of these ambiguities, sampling a wide range of districts remains a very difficult problem and the centre of mass of any district often gets locked in some metastable region. It is therefore difficult to obtain an ensemble that represent a wide range and often some particular map near the initial condition is overrepresented.

1.4 Computational Challenges

Many of the computational challenges present in sampling derive from complications arising from the complexities of the underlying energy functions and log-likelihoods. To estimate averages of the form (1.1), one needs a large number of samples such that the empirically obtained distribution is sufficiently close to the target. This is difficult

to achieve when systems get trapped in some specific region of high probability and never transitions to other regions, even though these other states may have similar probability mass. If the goal is to find the true average (1.1), this metastability will impede the convergence to the target distribution and the collected statistics will be biased.

In the Gaussian mixture example seen in the previous section the lack of exploration is not problematic as all regions of high probability are degenerate, due to the symmetry of the statistical model. However, an equivalent but much more complex real-world problem in MD is that of determining the stable configurational states of bio-molecules such as Alanine-12, illustrated in Figure 1.3. The helical configuration shown in this Figure is stable at 300K; but are there other configurations which are also stable at this temperature? This problem relates directly to the development of new medicines and could have real world consequences as other stable configurations are missed because of inferior sampling. In the coming chapters we will discuss the sampling of this molecule and show that to answer this question in full, we require accelerated sampling methods that speed up the convergence.

1.4.1 Long Term Stability and Accuracy

In Section 1.2 we introduced the concept of simulating a set of dynamical equations to propose configurations in phase-space that follow some probability distribution. The equations of motion describing such systems are discretised with a time step Δt and if we know the solution for some time $k\Delta t$ the solution at time $(k+1)\Delta t$ is found via some recurrence relation. We mentioned in the introduction that biological phenomena happen on the time-scale of milliseconds or seconds, and we denote the total length of the simulation as $\mathcal{T}$. This means that to complete a simulation of length $\mathcal{T} = 1ms$ we need to take K_{tot} number of steps, where,

$$\left\lceil \frac{\mathcal{T}}{\Delta t} \right\rceil = K_{\text{tot}} \in \mathbb{N}, \tag{1.38}$$

To illustrate that $K_{\text{tot}} \gg 1$ consider a deterministic particle described by the Hamiltonian,

$$H(q,p) = \frac{1}{2}p^2 + \frac{1}{2}\omega^2 q^2. \tag{1.39}$$

The potential describes a harmonic oscillator with a period $T = 2\pi/\omega$, with frequency $\nu = 1/T$. Using the deterministic dynamics given by (1.13), discretised by Strömer-Verlet [28], the recurrence relation stepping forward in time is equivalent to (see [7, sec.1.2.2.4]),

$$\begin{pmatrix} q_{k+1} \\ p_{k+1} \end{pmatrix} = A \begin{pmatrix} q_k \\ p_k \end{pmatrix}, \quad \text{where} \quad A = \begin{pmatrix} 1 - \frac{1}{2}(\omega\Delta t)^2 & \Delta t \\ -\omega\Delta t \left(1 - \frac{1}{4}(\omega\Delta t)^2\right) & 1 - \frac{1}{2}(\omega\Delta t)^2 \end{pmatrix}. \tag{1.40}$$

This dynamics is only stable if the eigenvalues of matrix A are smaller than 1. This is the case if and only if,

$$\Delta t < \frac{2}{\omega}. \tag{1.41}$$

This relation implies that the smallest conceivable timestep we may use for MD simulations is limited inversely by the order of the fastest oscillation in the system. These oscillations are commonly known to be chemical bonds to hydrogen atoms, roughly described as harmonic oscillators with a frequency of around 10^{15}s^{-1}. This simple

analysis therefore also translates into MD simulations, where the timestep is limited to the order of $\Delta t = 10^{-15}$s i.e. a femtosecond. From (1.38) it is clear that even for a millisecond length simulation, one needs $K_{\text{tot}} = 10^{12}$ i.e. on the order of a trillion steps.

As a result of this calculation we conclude that the numerical methods we must consider need to be stable in the long time limit, in which even small errors local in time become noticeable due to the large number of steps taken. The accumulation of errors on the invariant measure, is an active area of research and we refer the reader to [19, 29, 30], for papers on the subject.

For statistical models where one seeks to calculate the posterior distribution, the long-time limit is motivated by ergodic theory, which is discussed in Chapter 2. Heuristically this means that empirical averages only converge to space averages like (1.1) in the long time limit, such that a large number of steps always has to be evaluated to solve these types of problems. This common quality is how we motivate the close connection between MD simulation and Bayesian inference and sampling of more general statistical models.

1.4.2 Metastability

As we saw in the right panel of Figure 1.1, metastability of a dynamical system (or stochastic dynamical systems) is related to the presence of regions of high probability separated by regions of low probability. This issue also exists in MD where e.g the transition between any two configurational folded states of bio-molecules are rare events. In this situation the stable folded states are associated with high probability mass and the transition regions are represented by regions of low probability. These are the hallmarks for systems suffering from metastability.

Under these circumstances regions of high probability are oversampled as the system gets stuck in some region of phase space, sampling a particular configuration over and over. Similarly, regions of low probability mass are undersampled and the final result is heavily biased. This problem can also be understood in terms of transitions between regions of high probability mass, which are known as rare events if such transitions are unlikely to be observed.

In sizable systems of practical interest this becomes problematic as it is very difficult to assess whether a simulation has run for "long enough" and all rare events have been observed, such that the results can be considered to have converged. In these situations one is forced to work under the assumption that this is the case, even though certain knowledge that this is the case will never be obtained.

In this light, the design and development of methods that overcome these intrinsic limitations become important. More precisely, this is often quantified in terms of speed of convergence, where we seek to define methods that improve on the convergence speed of methods from a previous generation. This motivates our study of accelerated sampling schemes, where accelerated refers to an improved rate of convergence to the probability measure (or, to be more precise, improvement in the rate of convergence of observable averages).

Conceptually, one often consider metastability as the result of two types of phenomena: *energetic barriers* and *entropic barriers*. In the left panel of Figure 1.4 we illustrate a potential landscape with 4 metastable regions separated by energetic barriers. In particular, and because we use it for future illustrations, we state the potential

used to illustrate energetic barriers explicity as,

$$U(x_0, x_1) = \sum_{i=0}^{1} (x_i^2 - 1)^2 + (x_i + 1). \tag{1.42}$$

This potential is an uneven double well in two dimensions, which consequently has four metastable regions: lower left is the deepest well, on the diagonal are two wells of equivalent depths and the top right well is shallow. Below we use this landscape to illustrate the performance of both accelerated and standard sampling schemes. Metastability resulting from energetic barriers in the potential landscape can be often be overcome by increasing the temperature, as it in this case will directly increase the probability of transitioning from one region to another.

On the other hand, metastability resulting from entropic barriers does not respond to the temperature in this way. This type of barrier is illustrated to the right in Figure 1.4. These barriers are characterised by the slow diffusion along some reaction coordinate, which in the case of Figure 1.4 would be parallel to the line connecting the metastable regions. Often, to increase the exploration of this type of metastable landscape we require the knowledge of the slow reaction coordinate or learning it on-the-fly. We do not address this problem and instead focus our attention on metastabillity arising primarily as a result of energetic barriers. It should be noted that in real world systems the presence of metastability is often the consequence of both energetic and entropic barriers and it is then interesting to what extent a temperature-based scheme can enhance transitions. It is also intriguing to consider generalization of the temperature-based apparatus whereby an alternative collective variable is used as the mechanism of acceleration.

1.5 Original Contributions

Below we list the original contributions included in Chapters 4, 5 and 6 of this thesis.

Infinite Switch Simulated Tempering with Adaptive Weight Learning

This was joint work with Ben Leimkuhler, Jianfeng Lu and Eric Vanden-Eijnden and has been published in [31].

We extend the understanding of the simulated tempering method and show that equations that are swapping free can be implemented in the infinite switch limit. Following an argument similar to Dupuis et. al [32] we show that this limit is optimal. An inherent limitation of using the simulated tempering method is that it requires the knowledge of the partition function. We address this issue and perform a numerical study on a few simple examples and the Alanine-12 bio-molecule. An additional new section, not included in the original paper, briefly outlining the use of the algorithm for Gerrymandering has also been added.

Remarks: An implementation of this algorithm is available to use with all major molecular dynamics packages through MIST, Bethune et. al [33]. Note also that the introductory sections of this chapter, excluding the numerical experiment section, were rewritten for this thesis.

Isobaric-Isothermal Sampling

This was joint work with Ben Leimkuhler and Eric Vanden-Eijnden.

We derive a set of Lie-Trotter based isobaric-isothermal sampling schemes based on Langevin dynamics with a symplectic discretisation of the effective Hamiltonian. We discuss the implementation and discretisation in detail to aid the understanding of practitioners and we address a few abnormalities that has lead to some misunderstanding in the past. The chapter is finished with a set of numerical experiments illustrating the properties of the derived schemes. We make a suggestion based on these results for a general purpose barostat algorithm, that can used as the basis for future sampling schemes.

Remarks: This is unpublished work which continues with Brian Laird. Since the algorithm requires a special virial-type term, implementation in standard molecular dynamics packages is complex as the force loop has to be modified to efficently calculate it. Because of this, a significant amount of work has gone into developing all aspects of the algorithm (linked-cell, periodic cell, force fields, etc.), including the parallelism of a `C++` library. This will be released under GPL-3.

Generalised Simulated Tempering

This was joint work with Ben Leimkuhler.

In this exploratory final chapter, we investigate computationally the limitations and possibilities of the infinite switch sampling schemes of previous chapters. We demonstrate the limitations of the method and the importance of understanding the form of the collective variable to be tempered and illustrate cases when the algorithm does not perform well.

Remarks: This is unpublished work which was undertaken to investigate the boundaries of simulated tempering methods. The aim was to also to identify potential areas of future research.

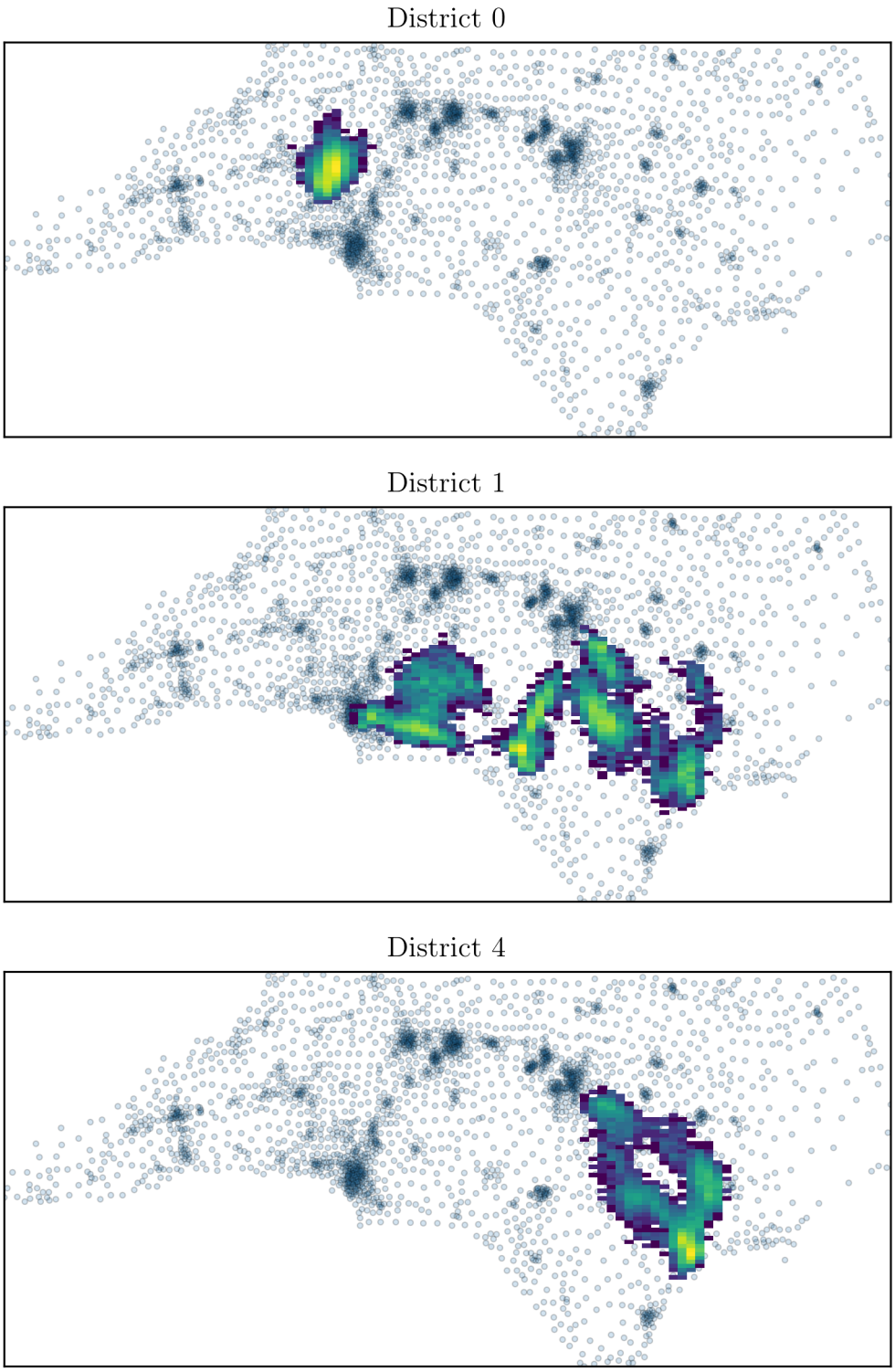

Figure 1.2: Plots of representative heatmaps of the centre of mass for three political districts overlaid on a scatter plot of the precincts in North Carolina. A lighter colour indicates a higher frequency, these changes in colour indicates that each district is metastable at the fictitious temperature $\beta = 5$. The samples were obtained using the Gerrymandering codebase from Duke university and uses an algorithm based on Metropolis hastings. See Section 4.3.4 for a brief outline of the accelerated sampling attempts for this model.

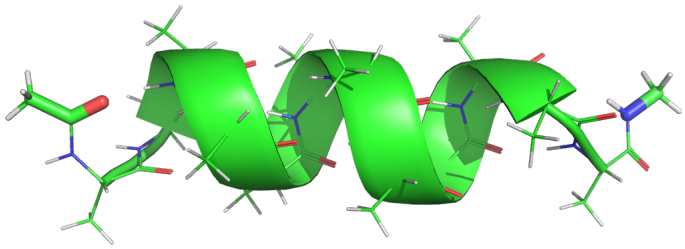

Figure 1.3: An illustration of the Alanine-12 bio molecule in a Helical configuration is shown and is stable at 300K.

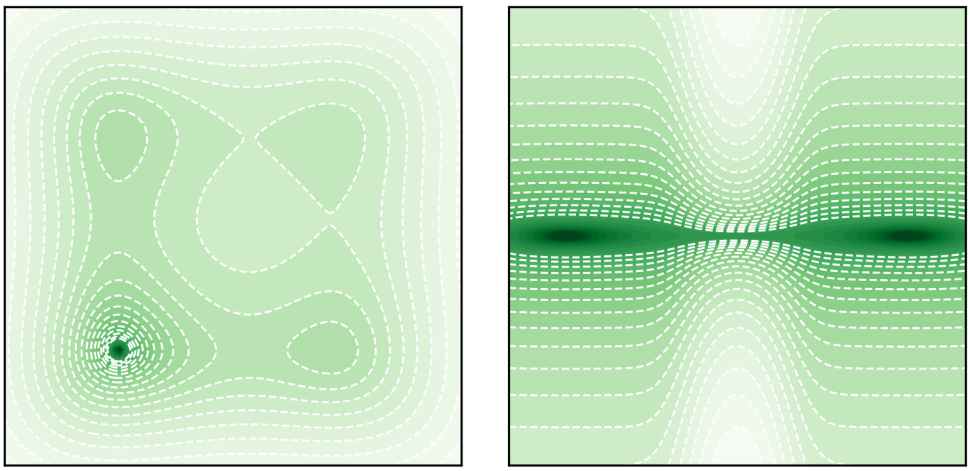

Figure 1.4: *Left.* This panel shows a metastable landscape with four basins separated by energetic barriers. These basins are of three different depths, where the deep and shallow basins are separated by two wells of identical depth. *Right.* This panel shows an entropic barrier between two metastable states.

Chapter 2

Foundations of Stochastic Numerical Methods

The aim of this chapter is to introduce several theoretical results that formalise important properties of sampling methods. We avoid the presentation of detailed proofs, which can be found within the numerous citations, and focus our discussion instead on the implications of these results. Naturally, we also include definitions that form the basis of the vocabulary used in later chapters. This section heavily relies on Lelièvre and Stoltz [7] and Pavliotis [18].

2.1 Foundations of Stochastic Processes

Sampling is currently one of the most popular approaches to solving high dimensional integrals of the form (1.1). A large number of sampling methods have been developed over many years going back to 1953 and the work of Metropolis et al. [34]. With the increase in computational power in recent years sampling has become a viable approach to solve a large number of problems in modern computational science.

Intrinsically the concept is very simple. However, it is founded on rigorous theoretical principles that characterise sampling methods and the properties they must possess. One such important concept is ergodicity, which is best summarised as "time averages converge to space averages". We refer to [18, 7] for an in depth treatment of this topic.

Formally the sampling challenge can be distilled to computations of the following form:

$$\int_{\Omega\times\mathbb{R}^n} \varphi\left(\boldsymbol{q},\boldsymbol{p}\right)\mu(\mathrm{d}\boldsymbol{q},\mathrm{d}\boldsymbol{p}) = \lim_{\mathcal{T}\to\infty}\frac{1}{\mathcal{T}}\int_0^{\mathcal{T}}\varphi\left(\boldsymbol{q}(t),\boldsymbol{p}(t)\right)\mathrm{d}t, \tag{2.1}$$

for some $\boldsymbol{q}(t)$ and $\boldsymbol{p}(t)$ following some stochastic process. Here, μ is a probability distribution and it states that the average of the observable φ under μ can be estimated by evaluating it along some trajectory $(\boldsymbol{q}(t),\boldsymbol{p}(t))$, whose dynamics preserves μ in the infinite time limit. This equality summarises precisely what is exploited in sampling, and it also makes the natural connection with dynamical systems apparent.

In general this suggests that, by recording a chain of samples from some probability distribution μ and collecting the empirical average for some observable φ along this chain, one obtains $\mathbb{E}_\mu\left[\varphi\right]$. Consequently, in discrete time simulations we use the

empirical estimate of (2.1) and write,

$$\mathbb{E}_{\mu}\left[\varphi\right] = \lim_{k\to\infty} \frac{1}{k} \sum_{i=0}^{k-1} \varphi\left(\boldsymbol{q}_i, \boldsymbol{p}_i\right). \tag{2.2}$$

The rigorous foundations of sampling schemes used to numerically estimate the sum in (2.2) with a finite length trajectory, are based on the theory of stochastic differential equations (SDEs), Markov chains (MC) and Markov jump processes (MJP). The relevant theory is introduced in Section 2.1.1 as statements regarding stochastic processes, invariant distributions and infinitesimal operators. In Section 2.1.2 we introduce the concept of ergodicity and in Section 2.1.3 we discuss the errors associated with numerical schemes estimating (2.2) and we then introduce basic sampling schemes in Section 2.2 that form the basis of most of the currently used methods.

2.1.1 Stochastic Processes

Let the sample space Ω be endowed with the σ-algebra $\mathcal{F}$, where the pair $(\Omega, \mathcal{F})$ is a measurable space. Let μ be a probability measure, also known as the law of the process, such that for a process X, we have

$$\mu\left(B\right) = \mathbb{P}\left[X^{-1}\left(B\right)\right] = \mathbb{P}\left(\omega \in \Omega; X(\omega) \in B\right), \quad B \in \mathcal{B}\left(\mathbb{R}^n\right). \tag{2.3}$$

The triplet $(\Omega, \mathcal{F}, \mathbb{P})$ is called a probability space. If the measure μ has Radon-Nikodym derivative with respect to the Lebesgue measure, we define the probability density function $\rho\left(x\right)$ as,

$$\mathrm{d}\mu\left(x\right) = \rho(x)\,\mathrm{d}x. \tag{2.4}$$

Here, $\rho(x) = \mathrm{d}\mu/\,\mathrm{d}x$ is exactly the Radon-Nikodym derivative and if it is well defined we say that the measure $\mathrm{d}\mu$ is absolutely continuous w.r.t. to the measure $\mathrm{d}x$. The short hand notation for denoting this is often written as $\mathrm{d}\mu \ll \mathrm{d}x$.

There are several important types of random variables that are used thorough science to study different types of phenomena. A relevant example for this thesis is the Gaussian or normal random variable denoted as $\mathcal{N}\left(m, \sigma^2\right)$, in which m refers to the mean and σ to the standard deviation.

Many different types of stochastic processes are studied in engineering and science for phenomena that are affected by random fluctuations, often referred to as noise. Such processes could in general be arbitrarily complicated to the point where it becomes troublesome to even generate them. Often one is forced to make assumptions on the nature of such processes to generate simplified models that are computationally feasible. For our purposes, an important and popular example is *Brownian motion*, named after the Scottish botanist Robert Brown who described it in 1827 whilst studying the trajectories of pollen grains suspended in water.

Let $\mathcal{T}$ be an ordered set which is either $\mathbb{R}^+ = [0, \infty)$ or $\mathbb{Z}^+ = \{0, 1, 2, \ldots\}$.

Definition 1 (Brownian Motion)**.** *A Brownian process $W(t) : \mathcal{T} \to \mathbb{R}$ is a stochastic process with the following properties:*

1. $W(0) = 0$
2. *$W(t)$ has independent increments on non-overlapping intervals i.e. for $t_1 < t_2 < \ldots t_{n+1}$, $W_{t_2} - W_{t_1}, \ldots, W_{t_{n+1}} - W_{t_n}, \ldots$ are independent.*

3. *If* $t > s \geq 0$ *then* $W(t) - W(s)$ *has a Gaussian distribution with mean* 0*, variance* $t - s$ *and density* $g_{(t-s),0}(x)$*.*

From this definition one deduces that $t \mapsto W(t)$ is almost surely continuous, for $t \in \mathcal{T}$. Similarly, the n-dimensional Brownian motion $\boldsymbol{W}(t) : \mathcal{T} \to \mathbb{R}^n$ is a collection of n independent one dimensional Brownian motions.

Another important type of stochastic process is a *Markov process*, which is a type of process that assumes that: given the present, the future is independent of the past. This property ensures that the process has no memory, and although this might not be completely true for many applications, it often simplifies their theoretical treatment. In fact, all the numerical methods studied in this thesis are Markov processes or Markov chains as they are referred to in discrete time. We now proceed to describe this important topic in more detail.

Define a filtration of a measurable space $(\Omega, \mathcal{F})$ as a family $\{\mathcal{F}_t : t \in \mathcal{T}\}$ of increasing sub σ-algebras of $\mathcal{F}$ i.e. $\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}$ for $s \leq t$. Denote the filtration generated by the stochastic process X_t as,

$$\mathcal{F}_t^X := \sigma\left(X_s : s \leq t\right). \tag{2.5}$$

Note, that the stochastic process X_t is adapted to the filtration $\{\mathcal{F}_t\}$ if X_t is a $\mathcal{F}_t$-measurable function for all $t \in \mathcal{T}$.

Definition 2 (Markov Process)**.** *A stochastic process* X *on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *taking values in* $\mathbb{R}$ *with respect to the filtration* $\left\{\mathcal{F}_t^X\right\}$ *is a Markov process with respect to the filtration* $\mathcal{F}_t^X$ *if,*

$$\mathbb{P}\left[X_t \in B | \mathcal{F}_s^X\right] = \mathbb{P}\left[X_t \in B | X_s\right], \tag{2.6}$$

for all $t, s \in \mathcal{T}$ *with* $t \geq s$ *and* $B \in \mathcal{B}\left(\mathbb{R}^n\right)$*.*

Since we are considering a dynamical process we need to define how the average behaviour evolves in time. This is accessible through the definition of the transition function which is denoted as,

$$p\left(s, \boldsymbol{x}, t, B\right) = \mathbb{P}\left[\boldsymbol{X}_t \in B | \boldsymbol{X}_s = \boldsymbol{x}\right]. \tag{2.7}$$

This defines the probability that $\boldsymbol{X}_t$ is in the set B at time t starting at the deterministic position $\boldsymbol{x}$ at initial time s. The transition function satisfies the famous Chapman-Kolmogorov equation,

$$p\left(s, \boldsymbol{x}, t, B\right) = \int_{\mathbb{R}} p(u, \boldsymbol{y}, t, B) p(s, \boldsymbol{x}, u, \mathrm{d}\boldsymbol{y}) \quad \text{for any } s \leq u \leq t. \tag{2.8}$$

If this transition function only depends on the time difference, the process is called *homogeneous* in time and implies that $p(s, \boldsymbol{x}, t, B) = p(0, \boldsymbol{x}, t - s, B)$. This allows us to simplify the notation of (2.8) to,

$$p\left(s + t, \boldsymbol{x}, B\right) = \int_{\mathbb{R}} p(t, \boldsymbol{y}, B) p(s, \boldsymbol{x}, \mathrm{d}\boldsymbol{y}). \tag{2.9}$$

Finally, we introduce the family of operators $\{\mathcal{P}_t\}_{t \in \mathcal{T}}$ that describe the evolution of the expectation of some deterministic $\varphi \in \mathcal{C}_b(\mathbb{R}^n)$ i.e. continuous, measurable and bounded function,

$$\mathcal{P}_t \varphi(\boldsymbol{x}) := \int_{\mathbb{R}^n} \varphi(\boldsymbol{y}) p(t, \boldsymbol{x}, \mathrm{d}\boldsymbol{y}) = \mathbb{E}\left[\varphi(\boldsymbol{X}_t) | \boldsymbol{X}_0 = \boldsymbol{x}\right]. \tag{2.10}$$

For the moment assume that for $t \geq 0$, $\mathcal{P}_t : \mathcal{C}_b(\mathbb{R}^n) \to \mathcal{C}_b(\mathbb{R}^n)$ and continuous in time[1]. This family of operators on $\mathcal{C}_b(\mathbb{R}^n)$ satisfy the *semigroup property*,

$$\begin{aligned} \mathcal{P}_{t+s}\varphi(\boldsymbol{x}) &= \int_{\mathbb{R}^n} \varphi(\boldsymbol{y}) p(t+s, \boldsymbol{x}, \mathrm{d}\boldsymbol{y}), \\ &= \int_{\mathbb{R}^n} \mathcal{P}_t\varphi(\boldsymbol{z}) p(s, \boldsymbol{x}, \mathrm{d}\boldsymbol{z}), \\ &= \left(\mathcal{P}_t\mathcal{P}_s\right)\varphi(x), \end{aligned} \tag{2.11}$$

which is obtained through (2.8). Similarly, it is straightforward to see that from definition (2.10) we have $\mathcal{P}_0 = I$. Define a new set $\mathcal{D}(\mathcal{L}) \subset \mathcal{C}_b(\mathbb{R}^n)$ of functions $\varphi \in \mathcal{C}_b(\mathbb{R}^n)$ for which the following limit exists in $\mathcal{C}_b(\mathbb{R}^n)$,

$$\mathcal{L}\varphi = \lim_{t \to 0^+} \frac{\mathcal{P}_t\varphi - \varphi}{t}. \tag{2.12}$$

The set $\mathcal{D}(\mathcal{L})$ is called the domain of the operator $\mathcal{L}$.

Definition 3 (Infinitesimal Operator)**.** *The operator $\mathcal{L} : \mathcal{D}(\mathcal{L}) \to \mathcal{C}_b(\mathbb{R}^n)$ defined in* (2.12) *is called the infinitesimal operator and is known as the generator of the Markov process X_t.*

2.1.2 Ergodicity

The semi-group (2.10) associated with the time homogeneous Markov process $\{X_t\}$ can be shown to satisfy the backward Kolmgorov equation. Introduce the deterministic function,

$$u(\boldsymbol{x}, t) = \int_{\mathbb{R}^n} \varphi(\boldsymbol{y}) p(t, \boldsymbol{x}, \mathrm{d}\boldsymbol{y}), \tag{2.13}$$

for $\varphi(\boldsymbol{x}) \in \mathcal{C}_b(\mathbb{R}^n)$. Then the backward Kolomgorov equation is,

$$\partial_t u(\boldsymbol{x}, t) = \mathcal{L}u(\boldsymbol{x}, t). \tag{2.14}$$

If the process X_t satisfies an SDE of the form,

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{b}(\boldsymbol{X})\,\mathrm{d}t + \sigma(\boldsymbol{X})\,\mathrm{d}\boldsymbol{W}, \tag{2.15}$$

for $\boldsymbol{b}$ and σ smooth and continuous, then the infinitesimal generator is the second order differential operator formally acting on smooth functions defined as,

$$\mathcal{L} := b^i(\boldsymbol{x})\,\partial_i + \frac{1}{2}\Sigma^{ij}(\boldsymbol{x})\,\partial_i\partial_j. \tag{2.16}$$

See e.g [18, theorem 2.7] and using Einstein notation to sum over repeated indices. Here, ∂_i denotes the partial derivative w.r.t. the i^{th} component and $\Sigma(\boldsymbol{x})$ is a matrix defined such that $\Sigma(\boldsymbol{x}) = \sigma(\boldsymbol{x})\sigma^T(\boldsymbol{x})$. In view of (2.14) it is customary to formally write the semi group operator as,

$$\mathcal{P}_t = \exp\left[t\mathcal{L}\right]. \tag{2.17}$$

[1]This property will be satisfied for all processes we consider.

Similarly, if the transition function $p(s, \boldsymbol{x}, \mathrm{d}\boldsymbol{y})$ has an absolutely continuous density $\rho(t, \boldsymbol{x}, \boldsymbol{y})$ w.r.t. to the Lebesgue measure, then by (2.13) we have,

$$\partial_t u(\boldsymbol{x}, t) = \int \varphi(\boldsymbol{y}) \partial_t \rho(t, \boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}. \tag{2.18}$$

Consider (2.14) and (2.13) by using integration by parts twice and Ito-calculus we have,

$$\partial_t u(t, \boldsymbol{x}) = \int \varphi(\boldsymbol{y}) \left(-\partial_i \left(b^i(\boldsymbol{y}) \rho(t, \boldsymbol{x}, \boldsymbol{y}) \right) + \frac{1}{2} \partial_i \partial_j \left(\Sigma^{ij}(\boldsymbol{y}) \rho(t, \boldsymbol{x}, \boldsymbol{y}) \right) \right) \mathrm{d}\boldsymbol{y}. \tag{2.19}$$

Equating (2.18) with (2.19), which is possible as a result of the time-homogeneity, the backward Kolomgorov eqution is written as

$$\partial_t \rho(t, \boldsymbol{x}, \boldsymbol{y}) = \mathcal{L}^\dagger \rho(t, \boldsymbol{x}, \boldsymbol{y}), \tag{2.20}$$

with the operator $\mathcal{L}^\dagger$ defined as,

$$\mathcal{L}^\dagger \rho(t, \boldsymbol{x}, \boldsymbol{y}) := -\partial_i \left(b^i(\boldsymbol{y}) \rho(t, \boldsymbol{x}, \boldsymbol{y}) \right) + \frac{1}{2} \partial_i \partial_j \left(\Sigma^{ij}(\boldsymbol{y}) \rho(t, \boldsymbol{x}, \boldsymbol{y}) \right). \tag{2.21}$$

The Forward Kolomgorov equation (2.20) will also be referred to as the Fokker-Planck equation.

Recall that the interest of this thesis lies with the computation of the long-term empirical average of some given observable for, in general, some time-homogeneous Markov processes. This means that we are interested in characterizing the long-term behaviour of such processes which takes its root in the concept of *ergodicity*. We define this as follows.

Definition 4 (Ergodicity)**.** *A Markov process $\{X_t\}$ is called ergodic if the equation*

$$\mathcal{L} g = 0, \quad g \in \mathcal{C}_b \left(\mathbb{R}^n \right), \tag{2.22}$$

has only constant solutions. Using (2.22) *and* (2.12) *this is equivalent to,*

$$\left(\mathcal{P}_t - I \right) g = 0, \tag{2.23}$$

only having constant solutions for $t > 0$.

Viewing these definitions in the light of (2.14) one deduces that these are expressions of vanishing time-derivatives in terms of space averages. This naturally leads us on to the concept of stationary, reversibility and irreducibly.

Definition 5 (Stationarity)**.** *A measure π is a stationary probability distribution for a Markov process with generator $\mathcal{L}$ if, for any smooth observable, φ,*

$$\int_\Omega \mathcal{L} \varphi(\boldsymbol{y}) \pi(\mathrm{d}\boldsymbol{y}) = 0. \tag{2.24}$$

This is equivalent to stating that the distribution of the process is invariant under time shift i.e. if $x = X_0 \sim \pi$ then $X_t \sim \pi$ for any time $t > 0$.

Definition 6 (Reversibility)**.** *Let φ_1 and φ_2 be two observables, $\{X_t\}$ is said to be reversible w.r.t. the probability distribution π, if the detailed balance condition is satisfied,*

$$\int_\Omega \varphi_1(\boldsymbol{y}) \left(\mathcal{L} \varphi_2 \right)(\boldsymbol{y}) \pi(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) = \int_\Omega \varphi_2(\boldsymbol{y}) \left(\mathcal{L} \varphi_1 \right)(\boldsymbol{y}) \pi(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}). \tag{2.25}$$

Note that, this condition implies stationary (consider $\varphi_1 = 1$). This property is very useful in practise and is explicitly used in the Metropolis Hastings algorithm, Section 2.2.1. Equality (2.25) *also states that if $\boldsymbol{x}_0 \sim \pi$ then $\{\boldsymbol{X}_t\}_{0<t\leq\mathcal{T}}$ has the same law as the time reversed path $\{\boldsymbol{X}_{T-t}\}_{0<t\leq\mathcal{T}}$ for all $T > 0$.*

Definition 7 (Irreducibility)**.** *A process $\{X_t\}$ is said to be irreducible if for any Borel set A with positive Lebesgue measure, and Lebesgue-almost all initial conditions $\boldsymbol{x} \in \mathbb{R}^n$ for all $t > 0$,*

$$\mathcal{P}_t\left(\mathbb{1}_A(\boldsymbol{x})\right) > 0. \tag{2.26}$$

This means that the process can reach a set of the state-space with positive probability in positive time starting from any initial condition. These properties characterise the conditions under which the transition function of the Markov process, $p(t, \boldsymbol{x}, \mathrm{d}\boldsymbol{y})$, admits a density. The existence of such measures are understood through the Hörmander condition which gives a criterion for obtaining the regularity of $p(t, \boldsymbol{x}, \mathrm{d}\boldsymbol{y})$. We say that the associated operator $\mathcal{L}$ of $p(t, \boldsymbol{x}, \mathrm{d}\boldsymbol{y})$ that satisfies Hörmander is *hypoelliptic*. To make this statement more precise, we introduce the Lie-algebra $L\left(A_0, \ldots, A_k\right)$ of a family of operators $(A_0, \ldots, A_k)$, which is the vector space containing the Span $(A_0, \ldots, A_k)$, i.e.

$$\text{If } B \in L\left(A_0, \ldots, A_k\right) \implies [B, A_i] \in L\left(A_0, \ldots, A_k\right). \tag{2.27}$$

Here the Lie bracket is defined as,

$$[A, B] := AB - BA. \tag{2.28}$$

Definition 8 (Hypoellipticity)**.** *Let $\mathcal{L}$ be an operator of a stochastic process and assume that we can express this operator as,*

$$\mathcal{L} = \frac{1}{2}\sum_{k=1}^{K} A_k^2 + A_0. \tag{2.29}$$

The Hörmander condition states for the operators $\{A_k\}_{k\geq 0}$ that,

$$L\left(A_0, \ldots A_K\right) = Span\left(\partial_{x_0}, \ldots, \partial_{x_n}\right). \tag{2.30}$$

If the Hörmander condition (2.30) *is satisfied then $\mathcal{L}$, $\mathcal{L}^\dagger$ and $\partial_t - \mathcal{L}^\dagger$ are all hypoelliptic. In particular, since $\partial_t - \mathcal{L}^\dagger$ is hypoelliptic: the law of the process, the transition function and the invariant measure all have densities.*

It can be shown that the process defined in (2.15) is hypoelliptic when $\boldsymbol{b}$ and σ are sufficiently smooth and σ has full rank. As a result their laws and transition functions have densities.

Theorem 1 (Sufficient Condition for Ergodicity)**.** *If* (2.30) *is satisfied then the law of the process admits a density w.r.t. Lebesgue measure, $\pi(\mathrm{d}\boldsymbol{y})$. Moreover, if an invariant measure exists and the process is irreducible then such an invariant measure is unique for $\varphi \in L^1(\pi)$ and,*

$$\lim_{\mathcal{T}\to 0} \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} \varphi(\boldsymbol{X}_t)\, \mathrm{d}t = \int_\Omega \varphi(\boldsymbol{y})\pi(\mathrm{d}\boldsymbol{y}) \quad a.s, \tag{2.31}$$

i.e. for any bounded measurable observable φ we have convergence in the sense of ergodic averages. If the process is also aperiodic then then it converges in law as well. This is assumed to be the case in the remainder of the thesis.

If the Hörmander condition holds, then noise acting in only a single direction is sufficient to affect uncoupled degrees of freedom such that the system diffuses in every direction. This gives us the existence of a density of the law and the transition probability, which we assume are always absolutely continuous w.r.t. Lebesgue.

2.1.3 Convergence and Error of Numerical Methods

In this section we discuss the errors associated with two types of implementations of stochastic numerical methods. Until now, we have only considered a stochastic process continuous both in space and time. It is possible to approximate such processes as a *Markov jump process* (see e.g [35, 36]). It is, however, far more common to consider methods that generate a Markov chain with no concept of time [34, 37] i.e. a method which generates a sequence of samples from some known distribution. These types of methods are often referred to as Markov chain monte carlo (MCMC) methods and the most basic of these are discussed in Section 2.2.1.

Another very popular approach to sampling is to consider time as a sequence of discrete increments Δt from some initial time t_0. These types of methods arise naturally when one considers a time discretisation of stochastic differential equations (SDEs), and will be the main focus of this thesis. Initially, basic results from Lie-Trotter splittings based on a symplectic splitting schemes of the Hamiltonian, are introduced in Section 2.2.2.

We can discuss both these approaches by considering a stochastic numerical method as a function

$$\Phi : \Theta \times \mathbb{R}^n \to \Theta, \tag{2.32}$$

where $\Theta = \Omega$ or $\Theta = \Omega \times \mathbb{R}^n$ depending on the order of the dynamics considered. This function combines the current state X_k with $Y \in \mathbb{R}^n$, an i.i.d random variable of the same dimensionality, to obtain the next step of the chain,

$$X_{k+1} = \Phi\left(X_k, Y\right). \tag{2.33}$$

Note that time discrete methods will be distinguished by the notation: $\Phi_{\Delta t}$. Consider the estimator of some bounded function $\varphi(X) : \mathbb{R}^n \to \mathbb{R}^n$ and denote it as,

$$\hat{\varphi}_N = \frac{1}{N} \sum_{k=1}^{N} \varphi\left(\boldsymbol{X}_k\right). \tag{2.34}$$

Assuming that the Markov chain is ergodic, our aim is to approximate the expectation of this observable as,

$$\lim_{N \gg 1} \hat{\varphi}_N \approx \int_{\Omega} \varphi(\boldsymbol{y}) \hat{\mu}(\mathrm{d}\boldsymbol{y}) = \mathbb{E}_{\hat{\mu}}\left[\varphi\right]. \tag{2.35}$$

Here, $\hat{\mu}\left(\mathrm{d}\boldsymbol{y}\right)$ is the invariant probability distribution of the Markov chain and will depend on the numerical scheme. We also introduce the exact probability distribution, $\mu\left(\mathrm{d}\boldsymbol{y}\right)$ and refer to it as the exact or target probability measure, which is the actual distribution that we want to sample. Note that (2.35) is also biased in the $N \to \infty$ limit i.e. $\hat{\mu} \neq \mu$ which is a result of the Δt bias which we make explicit next.

The exact quantity that we are interested in is denoted as $\mathbb{E}_{\mu}\left[\varphi\right]$, which is the observable which we seek to minimize errors with respect to. This means that we want to construct a method that generates samples from $\hat{\mu}$ which is a measure that is close, in some sense, to the exact μ. To make this statement more precise, we consider the

following equality,

$$\mathbb{E}\left[\left|\hat{\varphi}_N - \mathbb{E}_\mu\left[\hat{\varphi}\right]\right|^2\right] = \left(\mathbb{E}\left[\varphi_N\right] - \mathbb{E}_\mu\left[\varphi\right]\right)^2 + \mathbb{E}\left[\left|\hat{\varphi}_N - \mathbb{E}\left[\hat{\varphi}_N\right]\right|^2\right]. \tag{2.36}$$

Here, the two terms on the right hand side are the square of the bias and the square of the statistical error respectively. The statistical error is an expression of the central limit theorem arising from the law of large numbers in (2.35). Given that we have N Markov chain steps, this term will converge slowly with order $\mathcal{O}\left(N^{-1/2}\right)$. Depending on the numerical method we consider, generally speaking, the statistical error will dominate the bias.

To investigate the bias we expand it in two terms, finite sampling bias and perfect sampling bias,

$$\left|\mathbb{E}\left[\hat{\varphi}_N\right] - \mathbb{E}_\mu\left[\varphi\right]\right| \leq \left|\mathbb{E}\left[\hat{\varphi}_N\right] - \mathbb{E}_{\hat{\mu}}\left[\varphi\right]\right| + \left|\mathbb{E}_{\hat{\mu}}\left[\varphi\right] - \mathbb{E}_\mu\left[\varphi\right]\right|. \tag{2.37}$$

The first term is the finite sampling bias and measures the difference in expectation resulting from the use of a finite trajectory in the estimator and the expectation w.r.t. the numerical measure $\hat{\mu}$. The second term is the exact sampling bias and would be zero for a scheme with $\hat{\mu} = \mu$. This is precisely the case for the Metropolis-Hastings algorithm. In contrast, for time discrete schemes the numerical measure $\hat{\mu}$ is different from the exact measure μ and this term is no longer zero. Consequently, to study time-discrete methods one needs to control the imperfect sampling bias that these methods emit. This is often expressed as bounds of the form [38],

$$\left|\mathbb{E}_{\hat{\mu}}\left[\varphi(\boldsymbol{x})\right] - \mathbb{E}_\mu\left[\varphi(\boldsymbol{x})\right]\right| \leq C\Delta t^p, \tag{2.38}$$

and we say that the method is of order p of the invariant measure (large or equal to weak order). This automatically induces a measure under which we compare different numerical schemes and we favour methods of higher order – given a fixed computational cost. Therefore, if we wish to construct numerical schemes we can outline the rough goal as: construct time-discrete methods with a high order of accuracy, whilst keeping the computational cost small. In the MD setting, this means minimizing the number of force evaluations per timestep. Unfortunately, it is not entirely this straightforward in practice.

The constant in (2.38) may differ between two methods of the same order, which is often very difficult to estimate analytically and often only accessible via numerical experiments. In Section 2.2.2 we present an overview of a few relevant design considerations to make in the development of schemes for MD and establish conditions under which these methods are ergodic.

2.2 Methods for High Dimensional Sampling

In this section we present techniques used in practical situations to generate Markov-chains that sample some known distribution. These methods are basic and will only perform well under quite restrictive conditions, such as when the log-likelihood of the invariant probability density is not multimodal. To overcome these issues one needs to employ accelerated methods which will be addressed in Chapter 3. However, it is useful to consider the construction of simple methods on which the improved schemes are based, as it illustrates many of the important aspects in the construction of sampling schemes.

Following the theme of this thesis, we introduce the following schemes in the light of canonical sampling for MD – that is – constant temperature sampling. This helps emphasise how these techniques are used in practise, whilst also introducing basic concepts useful for understanding the accelerated sampling schemes.

2.2.1 Metropolis Hastings Algorithm

Below we utilize the concepts introduced in 2.1.1 and 2.1.2 to review some of the most popular methods commonly used to solve sampling problems numerically. One such fundamental example was first introduced in 1953 by Metropolis et. al. [34] and later generalized by Hastings in the 70's [37]. This method was consequently named the *Metropolis-Hastings Algorithm* and forms the basic structure for a vast array of methods that improve on the original algorithm. In fact an entire family of related methods are often referred to as *Markov chain Monte Carlo* (MCMC) methods.

In essence, these methods construct Markov chains with some desired invariant distribution of interest and are employed to empirically study integrals of the form (1.1). The methods therefore fall under the category of numerical integration schemes and could, for that reason, be compared to *quadrature* methods. The field of quadrature schemes is in itself rich, but quadrature methods are often intrinsically limited by the dimension of the integral, with standard quadrature schemes performing very well up to around $\mathbb{R}^3$. Although an improvement on quadrature methods, the large number of available sampling methods in existence, should indicate to the reader that this approach is not without its own complications and limitations. Some of the general issues plaguing sampling were touched upon in Section 1.3.

Let us return to the problem at hand. Assume that we are interested in calculating an integral of the form,

$$\mathbb{E}_\mu [\varphi] = \int_\Omega \varphi(\boldsymbol{q}) \mu (\mathrm{d}\boldsymbol{q}) , \tag{2.39}$$

which is the expectation of some observable φ under some probability measure $\mu(\mathrm{d}\boldsymbol{q})$. Interpreting this as a sampling problem, the goal is to construct a method which generates a Markov-Chain with invariant distribution,

$$\mu (\mathrm{d}\boldsymbol{q}) = Z^{-1} \rho(\boldsymbol{q}) \, \mathrm{d}\boldsymbol{q}, \quad \text{where} \quad Z := \int_\Omega \rho(\boldsymbol{q}) \, \mathrm{d}\boldsymbol{q}. \tag{2.40}$$

Here, we have introduced the normalisation constant Z which assures that $\int \mu = 1$ such that $\rho(\boldsymbol{q})$ is a probability density. The Metropolis-Hastings algorithm can be used to generate a Markov chain whose invariant probability density is $\rho(\boldsymbol{q})$. This is achieved in two steps, first a proposal step $\tilde{\boldsymbol{q}}$ is generated according to some transition kernel $T(\boldsymbol{q}, \mathrm{d}\tilde{\boldsymbol{q}})$ from the current position $\boldsymbol{q}$. Then the proposed step $\tilde{\boldsymbol{q}}$ is accepted or rejected with some probability. This is explicitly summarised in the following steps: starting from a given $\boldsymbol{q}_0$, and for $i \geq 1$,

1. Generate a proposal $\tilde{\boldsymbol{q}}$ from the transition kernel density,

$$\tilde{\boldsymbol{q}} \sim T (\boldsymbol{q}_i, \cdot) . \tag{2.41}$$

   Define the ratio $r(\boldsymbol{q}_i, \tilde{\boldsymbol{q}})$ as,

$$r(\boldsymbol{q}_i, \tilde{\boldsymbol{q}}) = \frac{\mu(\mathrm{d}\tilde{\boldsymbol{q}}) \; T (\tilde{\boldsymbol{q}}, \mathrm{d}\boldsymbol{q}_i)}{\mu(\mathrm{d}\boldsymbol{q}_i) \; T (\boldsymbol{q}_i, \mathrm{d}\tilde{\boldsymbol{q}})}. \tag{2.42}$$

2. To satisfy reversibility (detailed balance), we accept the proposal (i.e. set $\boldsymbol{q}_{i+1} = \tilde{\boldsymbol{q}}$) with probability

$$\min\left(1, r\left(\boldsymbol{q}_i, \tilde{\boldsymbol{q}}\right)\right), \tag{2.43}$$

otherwise let $\boldsymbol{q}_{i+1} = \boldsymbol{q}_i$.

It is clear from equation (2.43) that the acceptance calculation is independent of the normalisation constant Z in (2.40)[2]. This is an important quality of the method as the calculation of Z is in general a very difficult problem which the algorithm avoids.

To exercise control over the proposal move, one must choose an appropriate transition kernel. In general, a sensible choice would be a transition kernel which is similar to the invariant measure $\mu(\mathrm{d}\boldsymbol{q})$, such that proposals in regions of high probability are more likely. In general $\mu(\mathrm{d}\boldsymbol{q})$ is of course not known and choosing an appropriate transition kernel is difficult. In an abstract sense the choice reduces to a symmetric or non-symmetric kernel.

Symmetric transition kernels are the simplest type to implement and it was also the original type of kernel introduced by Metropolis [34]. If a transition kernel satisfies,

$$T(\boldsymbol{q}, \mathrm{d}\tilde{\boldsymbol{q}})\,\mathrm{d}\boldsymbol{q} = T(\tilde{\boldsymbol{q}}, \mathrm{d}\boldsymbol{q})\,\mathrm{d}\tilde{\boldsymbol{q}}, \tag{2.44}$$

it is said to be symmetric and (2.42) is independent of the normalisation constant for $T\left(\boldsymbol{q}, \cdot\right)$. A non-symmetric transition kernel does not satisfy this condition and can be used to bias the proposal towards regions of high probability, by making certain moves more likely. A standard example of this type is the Metropolis-Adjusted Langevin algorithm (MALA) [39, 40], which uses the gradient information, $\nabla_{\boldsymbol{q}} \log p(\boldsymbol{q})$, to bias proposals towards regions of high probability.

In general, any transition kernel must result in a positive and well defined ratio (2.42), which leads to the condition that $\mu(\mathrm{d}\tilde{\boldsymbol{q}})T\left(\tilde{\boldsymbol{q}}, \mathrm{d}\boldsymbol{q}_i\right)$ and $\mu(\mathrm{d}\boldsymbol{q}_i)T\left(\boldsymbol{q}_i, \mathrm{d}\tilde{\boldsymbol{q}}\right)$ must be mutually absolutely continuous for all $\boldsymbol{q}$ and $\tilde{\boldsymbol{q}}$. Additionally, it is straightforward to show that under these conditions the Metropolis-Hastings algorithm produces a Markov chain that satisfies detailed balance such that $\mu(\mathrm{d}\boldsymbol{q})$ is the invariant distribution [7].

Symmetric Transition Kernel: Random Walk

Let us consider a simple toy model for the simplest form of transition kernel based on a random walk. To this end, we use the uneven double well potential,

$$U(\boldsymbol{q}) = \sum_{i=1}^{2} 4(q_i^2 - 1)^2 + (q_i + 1). \tag{2.45}$$

Let the invariant distribution of interest be the configurational part of the canonical distribution for $\boldsymbol{q} \in \Omega = \mathbb{R}^n$,

$$\mu_\beta(\mathrm{d}\boldsymbol{q}) = e^{-\beta U(\boldsymbol{q})}\,\mathrm{d}\boldsymbol{q}. \tag{2.46}$$

To sample this distribution using Metropolis-Hastings, one must define a transition kernel. Let this transition kernel be such that the current position $\boldsymbol{q}$ is modified by the addition of a Gaussian perturbation. More precisely,

$$\tilde{\boldsymbol{q}} = \boldsymbol{q} + \boldsymbol{g}, \quad \boldsymbol{g} \sim \mathcal{N}\left(0, \sigma^2 I\right). \tag{2.47}$$

[2]Note that it may still depend on the normalisation constant of the transition kernel, T.

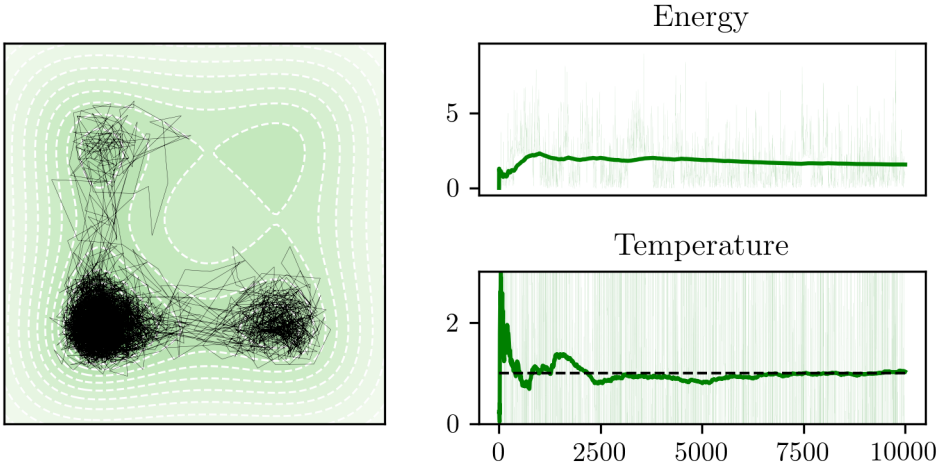

Figure 2.1: This figure shows a Metropolis Hastings trajectory $\sigma = 0.4$ and steps 10000. The long term average is indicated by the thick line and the instantaneous values are shown in the background. The expected temperature $T = 1$ in this case is shown as dashed black.

In this case, the transition kernel takes the form,

$$T(\boldsymbol{q}, \mathrm{d}\tilde{\boldsymbol{q}}) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{|\boldsymbol{q} - \tilde{\boldsymbol{q}}|^2}{2\sigma^2}\right] \mathrm{d}\tilde{\boldsymbol{q}}, \tag{2.48}$$

which is a symmetric proposal.

In the left panel of Figure 2.1 a trajectory of a Metropolis-Hastings trajectory is shown for $\sigma = 0.4$, $\beta = 1$ and $n = 2$. From the black trajectory, we see how the algorithm produces samples from (2.46), where (2.45) is shown as a green contour plot. The potential (2.45) illustrates the meta-stability issue and shows how the algorithm over samples regions of high probability whilst not exploring the full landscape. In fact for this finite time trajectory it completely avoids the entire high energy meta stable state in the top right hand corner.

This behaviour can be alleviated by tuning the σ parameter but determining the optimal choice of σ is difficult for a general potential $U(\boldsymbol{q})$. However, theoretical predictions exist for special cases and are determined in terms of the optimal acceptance rate – the ratio of accepted to rejected proposals. One such special case is the n dimensional Gaussian or Harmonic well if referring to the potential, and it was established by Roberts et. al. [40] that for this special case, the optimal acceptance rate should be 0.234. In more general cases no such estimate exists and σ should be picked to balance accepting a new step with exploring the potential to produce a sufficient number of non-correlated samples.

2.2.2 Continuous Stochastic Dynamics

Molecular Dynamics is in its most fundamental form an implementation of Newton's equations of motion and as such derives from the study of Hamiltonian systems. Numerical implementations of Hamiltonian dynamics are at this point well understood, and properties of algorithms have been studied in great detail. We refer to Leimkuhler

and Reich [12] or Hairer et al. [13] and references within for detailed studies of the properties of these schemes.

The standard algorithm for integrating Hamiltonian systems is known as Störmer-Verlet and was popularised for Molecular Dynamics by Verlet in 1967 [28]. This algorithm preserves energy, is symplectic (preserves integrals over phase-space volume) and is, as a result, stable in the long-time limit. This is precisely the limit relevant to MD and is the reason why the algorithm still is the standard choice in many MD packages. As we discuss below, these properties also have some relevance in the discretisation of SDEs in terms of volume preservation, see Abdulle et al. [17].

We mentioned in the introduction that an exact Hamiltonian flow preserves energy, which also holds true for the Störmer-Verlet scheme. When sampling constant temperature-volume conditions of particle systems, we are not concerned with the preservation of the energy; but we are instead interested in the conservation of the canonical measure. With this aim in mind, Nosé [41] and Hoover [42] introduced perturbed Hamiltonian dynamics, which preserves the canonical measure. However, this dynamics has been proved to not be ergodic for the harmonic oscillator [43] and as such we only mention it for historical completeness.

Fully ergodic dynamics and physically more relevant equations of motion, can be found in Langevin dynamics. The formal derivation of these equations was briefly outlined in the introduction and we repeat that: heuristically, Langevin dynamics describe large particles that exchange energy with a heat bath of smaller particles. We only consider the simplest case of generalised Langevin equation where the friction is a scalar coefficient γ. We write,

$$\begin{aligned} \mathrm{d}\boldsymbol{q} &= M^{-1}\boldsymbol{p}, \\ \mathrm{d}\boldsymbol{p} &= -\nabla_{\boldsymbol{q}} U(\boldsymbol{q})\,\mathrm{d}t - \gamma M^{-1}\boldsymbol{p}\,\mathrm{d}t + \sigma\,\mathrm{d}\boldsymbol{W}. \end{aligned} \tag{2.49}$$

In this case, the fluctuation matrix term σ is related to the friction term via the fluctuation-dissipation relation, which for the case at hand reads,

$$\sigma\sigma^T = \frac{2\gamma}{\beta} I, \tag{2.50}$$

where I is the appropriate identity matrix. As discussed in the introduction, this ensures that the invariant measure of (2.49) is,

$$\mu(\mathrm{d}\boldsymbol{q}, \mathrm{d}\boldsymbol{p}) = \exp\left[-\beta H\left(\boldsymbol{q}, \boldsymbol{p}\right)\right] \mathrm{d}\boldsymbol{q}\,\mathrm{d}\boldsymbol{p}. \tag{2.51}$$

To verify that this is true we write down the generator of the Langevin process as the sum of two generators,

$$\mathcal{L} = \mathcal{L}_{\mathrm{H}} + \mathcal{L}_{\mathrm{OU}}. \tag{2.52}$$

The first term is the associated generator of Hamiltonian dynamics and takes the form,

$$\mathcal{L}_{\mathrm{H}}\varphi = M^{-1}\boldsymbol{p} \cdot \nabla_{\boldsymbol{q}}\varphi - \nabla_{\boldsymbol{q}} U\left(\boldsymbol{q}\right) \cdot \nabla_{\boldsymbol{p}}\varphi. \tag{2.53}$$

The second term is the generator of the *Ornstein-Uhlenbeck* (OU) process and takes the form,

$$\begin{aligned} \mathcal{L}_{\mathrm{OU}}\varphi &= -\gamma M^{-1}\boldsymbol{p} \cdot \nabla_{\boldsymbol{p}}\varphi + \beta^{-1}\gamma \nabla_{\boldsymbol{p}} \cdot \nabla_{\boldsymbol{p}}\varphi, \\ &= \gamma\beta^{-1} e^{\beta H(\boldsymbol{q},\boldsymbol{p})} \nabla_{\boldsymbol{p}} \cdot \left(e^{-\beta H(\boldsymbol{q},\boldsymbol{p})} \nabla_{\boldsymbol{p}}\varphi\right). \end{aligned} \tag{2.54}$$

With these equalities we show that the evolution of some observable $\varphi(\boldsymbol{q},\boldsymbol{p})$ under the evolution of the process (2.49) is invariant w.r.t. (2.51),

$$\begin{aligned}\int_{\Omega\times\mathbb{R}^n} &\mathcal{L}(\varphi)\,\mu(\mathrm{d}\boldsymbol{q},\mathrm{d}\boldsymbol{p}),\\ &= \int_{\Omega\times\mathbb{R}^n} \left(\mathcal{L}_{\mathrm{H}}(\varphi)+\mathcal{L}_{\mathrm{OU}}(\varphi)\right) e^{-\beta H(\boldsymbol{q},\boldsymbol{p})}\,\mathrm{d}\boldsymbol{q}\,\mathrm{d}\boldsymbol{p},\\ &= -\beta^{-1}\int_{\Omega\times\mathbb{R}^n} \nabla_{\boldsymbol{p}} e^{-\beta H}\cdot\nabla_{\boldsymbol{q}}\varphi - \nabla_{\boldsymbol{q}} e^{-\beta H}\cdot\nabla_{\boldsymbol{p}}\varphi\,\mathrm{d}\boldsymbol{q}\,\mathrm{d}\boldsymbol{p},\\ &+ \beta^{-1}\gamma\int_{\Omega\times\mathbb{R}^n} \nabla_{\boldsymbol{p}}\cdot\left(e^{-\beta H(\boldsymbol{q},\boldsymbol{p})}\nabla_{\boldsymbol{p}}\varphi\right)\mathrm{d}\boldsymbol{q}\,\mathrm{d}\boldsymbol{p},\\ &= 0.\end{aligned} \tag{2.55}$$

Both integrals are shown to be zero via application of integration of parts.

Under appropriate smoothness conditions on the potential $U(\boldsymbol{q})$, it can be shown that the Langevin process satisfies the Hörmander condition, and as a consequence the process emits a measure. This measure is uniquely defined as (2.51) , see [38, 14].

Next, we explain in more detail why it is beneficial to consider the operator of the Langevin process as being constructed from two sub generators: deterministic flow $\mathcal{L}_{\mathrm{H}}$ and stochastic diffusion $\mathcal{L}_{\mathrm{OU}}$. Note that in fact, the OU process can be solved exactly using an integration factor,

$$\begin{aligned}\mathrm{d}\boldsymbol{p} &= -\gamma M^{-1}\boldsymbol{p} + \sigma\,\mathrm{d}\boldsymbol{W},\\ \implies \boldsymbol{p}(t+\Delta t) &= e^{-\gamma M^{-1}\Delta t}\boldsymbol{p}(t) + \sigma\int_0^{\Delta t} e^{-\gamma M^{-1}(\Delta t-s)}\,\mathrm{d}\boldsymbol{W}(s).\end{aligned} \tag{2.56}$$

Using Itô calculus, the integral term has variance,

$$\begin{aligned}\mathbb{E}\left[\left(\int_0^{\Delta t} e^{-\gamma M^{-1}(\Delta t-s)}\,\mathrm{d}\boldsymbol{W}(s)\right)\left(\int_0^{\Delta t} e^{-\gamma M^{-1}(\Delta t-s)}\,\mathrm{d}\boldsymbol{W}(s)\right)^T\right]&\\ = \frac{1}{2\gamma}\left(I - e^{-2\gamma M^{-1}\Delta t}\right)M.&\end{aligned} \tag{2.57}$$

This implies that (2.56) is equal in law to,

$$\boldsymbol{p}(t+\Delta t) = e^{-\gamma M^{-1}\Delta t}\boldsymbol{p}(t) + \Sigma\boldsymbol{Y}_t, \tag{2.58}$$

where $\boldsymbol{Y}_t \sim \mathcal{N}(0,1)$ and $\Sigma\Sigma^T = \beta^{-1}\left(I - \exp[-2\gamma M^{-1}\Delta t]\right)M$. If we reconsider the interpretation of the fluctuation-dissipation theory and instead consider,

$$\sigma\sigma^T = 2\beta^{-1}\gamma M, \tag{2.59}$$

the friction is proportional to the momentum instead of the velocity. This has the effect of avoiding the calculation of the matrix exponential in (2.58), and the momentum is instead evaluated as,

$$\boldsymbol{p}(t+\Delta t) = e^{-\gamma\Delta t}\boldsymbol{p}(t) + \Sigma\boldsymbol{Y}_t, \tag{2.60}$$

where $\Sigma\Sigma^T = \beta^{-1}\left(I - \exp[-2\gamma\Delta t]\right)M$. The numerical evaluation of (2.58) is fine as long as the mass matrix is diagonal since the matrix exponential also is in this case. This does not hold true for NPT and so in that case it is beneficial to directly consider the friction proportional to the momentum instead of velocity.

Regardless of this interpretation, the Cholesky factorization of the mass matrix must be calculated, implying that it is always beneficial to work with a diagonal mass matrix.

In contrast to the OU process (2.54), a general solution to the Hamiltonian flow (2.53) cannot be evaluated analytically. A huge amount of literature exists on the design of numerical methods for these equations and the error that their discretisations emit [12, 13], with one such example being *Lie-Trotter* splitting schemes [44, 45] – a family that, for deterministic dynamics, contains the popular Störmer-Verlet scheme from earlier. A particularly well studied area is that of symplectic Lie-Trotter; methods that conserve the energy, the symplectic two-form $\mathrm{d}\boldsymbol{q}\wedge\mathrm{d}\boldsymbol{p}$ and therefore volume. We will focus our attention on stochastic Lie-Trotter schemes that split the Langevin process as in (2.52), with a symplectic integration of the Hamiltonian – treating the diffusive part (2.54) exactly.

For our purposes we follow the notation in [19] and denote the flow of these methods as,

$$\left[\boldsymbol{q}_{n+1}, \boldsymbol{p}_{n+1}\right]^T = \Phi_h \circ \Theta_{\Delta t,n}\left(\boldsymbol{q}, \boldsymbol{p}\right), \tag{2.61}$$

where $\Phi_{\Delta t}$ approximates the Hamiltoniain flow,

$$\begin{aligned} \mathrm{d}\boldsymbol{q} &= M^{-1}\boldsymbol{p}, \\ \mathrm{d}\boldsymbol{p} &= -\nabla U(\boldsymbol{q}). \end{aligned} \tag{2.62}$$

In (2.61) $\Theta_{\Delta t,n}$ denotes the exact flow in law of the OU process,

$$\begin{aligned} \mathrm{d}\boldsymbol{q} &= 0, \\ \mathrm{d}\boldsymbol{p} &= -\gamma M^{-1}\boldsymbol{p}\,\mathrm{d}t + \sigma\,\mathrm{d}\boldsymbol{W}. \end{aligned} \tag{2.63}$$

The ergodicity, statistical and numerical properties of this type of scheme was first studied in [46] where it was referred to as the *Geometric Langevin Algorithm*. In that paper the authors required $\Phi_{\Delta t}$ to be symplectic for the method to be ergodic, and also showed that the weak error of the full scheme is determined by the energy error of $\Phi_{\Delta t}$. This requirement is not necessary and it is possible to demonstrate that the dynamics is ergodic by showing that the numerical flow is irreducible, which was done shortly after for a set of higher-order methods [16, 47], that have become very popular in MD. In fact, it was shown by Abdulle et al. [17] that in general, the weak error is not determined by the order of $\Phi_{\Delta t}$, given that the flow is volume preserving. Indeed it was shown in that paper that if the energy error for the Hamiltonian flow is Δt^p, a numerical scheme with weak error Δt^{p+1} can be constructed.

This implies that symplecticity is a sufficient but not required property in order to capture the invariant measure of the Langevin process with high order. Although, through the application of these techniques one constructs numerical schemes of high order accuracy, as noted [19] the resulting methods might find limited use in applications such as MD, due to the requirement for multiple gradient evaluations per timestep.

A successful approach for the construction of numerical schemes that find use in MD, are those generated by the symmetric pattern introduced by Leimkuhler and Matthews [16]. Below we demonstrate the approach taken in the systematic design of these methods via the study of the global weak-error. Following this pattern it can be shown that under certain conditions, methods with a $\mathcal{O}\left(\Delta t^2\right)$ global weak error can be constructed, for the cost of a single gradient evaluation.

By working in the setting of weak backward analysis for SDEs [48, 15, 47, 19, 20],

our aim is to construct discretisations of (2.49) and (2.52) by cancellation of stochastic Taylor expansions in the timestep Δt. Following Zygalakis [15] the expectation of some observable is denoted as $u(t, \boldsymbol{q}, \boldsymbol{p}) = \mathbb{E}\left[\varphi(\boldsymbol{q}(t), \boldsymbol{p}(t)) | \boldsymbol{q}(0) = \boldsymbol{q}, \boldsymbol{p}(0) = \boldsymbol{p}\right]$. With the appropriate conditions on the gradient and noise, this satisfies the backward kolomogorov equation,

$$\partial_t u(t, \boldsymbol{q}, \boldsymbol{p}) = \mathcal{L} u(t, \boldsymbol{q}, \boldsymbol{p}). \tag{2.64}$$

Similar to ODEs, the local weak error can be used to characterise the global weak error and we say that a method of local weak error Δt^{p+1} has p global weak error, $\mathcal{O}\left(\Delta t^p\right)$. The goal is to express the error as a truncated series expansion in orders of Δt over a single timestep. This is then used to express the global weak error of the method. Integrating (2.64) over one timestep Δt,

$$u(\Delta t, \boldsymbol{q}, \boldsymbol{p}) - u(0, \boldsymbol{q}, \boldsymbol{p}) = \mathcal{L} \int_0^{\Delta t} u(s, \boldsymbol{q}, \boldsymbol{p}) \, \mathrm{d}s. \tag{2.65}$$

Using the stochastic Taylor expansion [49] we expand $u(s, \boldsymbol{q}, \boldsymbol{p})$ in time and obtain,

$$\begin{aligned} u(\Delta t, \boldsymbol{q}, \boldsymbol{p}) - \varphi(\boldsymbol{q}, \boldsymbol{p}) =& \Delta t \mathcal{L} \varphi(\boldsymbol{q}, \boldsymbol{p}) \\ &+ \sum_{k=1}^{N} \frac{\Delta t^{k+1}}{(k+1)!} \mathcal{L} \frac{\partial^k u(s, \boldsymbol{q}, \boldsymbol{p})}{\partial s^k} + \mathcal{O}\left(\Delta t^{N+2}\right). \end{aligned} \tag{2.66}$$

Recalling (2.64) this expansion is expressed as a single sum,

$$u(\Delta t, \boldsymbol{q}, \boldsymbol{p}) = \varphi(\boldsymbol{q}, \boldsymbol{p}) + \sum_{k=0}^{N} \frac{\Delta t^{k+1}}{(k+1)!} \mathcal{L}^{k+1} \varphi(\boldsymbol{q}, \boldsymbol{p}) + \mathcal{O}\left(\Delta t^{N+2}\right). \tag{2.67}$$

Note that (2.67) is a formal expansion that holds for all u continuous and differentiable up to arbitrary order.

To study the local error of a method for a single timestep, we assume that a discretisation of (2.49) gives the following expansion of some observable over the same step,

$$\mathcal{P}_{\Delta t} \varphi = \varphi + \Delta t \mathcal{A}_1 \varphi + \ldots + \frac{\Delta t^{p+1}}{(p+1)!} \mathcal{A}_{p+1} \varphi + \mathcal{O}\left(\Delta t^{p+2}\right), \tag{2.68}$$

where $\mathcal{A}_j$ for $0 \leq j < p+2$ are generators that depend on the choice of numerical scheme. A method for which $\mathcal{A}_1 = \mathcal{L}$ is called *consistent* and holds true for the methods we study. It is easy to see that for all consistent numerical schemes,

$$u\left(\Delta t, \boldsymbol{q}, \boldsymbol{p}\right) - \mathcal{P}_{\Delta t} \varphi = \mathcal{O}\left(\Delta t^2\right), \tag{2.69}$$

and the global weak error is said to be of order 1.

It is clear that a favourable scheme maximizes the number of cancellations in (2.69) for a minimal computational effort and to determine to which order this happens one must calculate the generators $\mathcal{A}_j$ in (2.68).

The methods which we present below were first introduced by Leimkuhler and Matthews [16] and use a symmetric construction based on the Störmer-Verlet scheme combined with an exact solve of the OU process as in GLA [46]. This implies that the Lie-Trotter schemes we consider from now on, for the Langevin process (2.49), are

based on the following constitutive generators,

$$\begin{aligned}\mathcal{L}_A\varphi &= M^{-1}\boldsymbol{p}\cdot\nabla_{\boldsymbol{q}}\varphi, \qquad \mathcal{L}_B\varphi = -\nabla_{\boldsymbol{q}}U\left(\boldsymbol{q}\right)\cdot\nabla_{\boldsymbol{p}}\varphi,\\ \mathcal{L}_O\varphi &= -\gamma\boldsymbol{p}\cdot\nabla_{\boldsymbol{p}}\varphi + \beta^{-1}\gamma\nabla_{\boldsymbol{p}}\cdot M\nabla_{\boldsymbol{p}}\varphi.\end{aligned} \tag{2.70}$$

Although it is possible to form schemes of arbitrary length, permuting over any number of combinations, we limit the discussion here to symmetric schemes of five letter length. As noted in [16] the terms in (2.68) can be found by utilising the Baker-Campbell-Hausdorff (BCH) formula [13] for expansions of operator exponentials. In particular we focus on the scheme labelled $BAOAB$, which gives the one step semi-group,

$$\mathcal{P}^{\mathrm{B}}_{\Delta t/2}\circ\mathcal{P}^{\mathrm{A}}_{\Delta t/2}\circ\mathcal{P}^{\mathrm{O}}_{\Delta t}\circ\mathcal{P}^{\mathrm{A}}_{\Delta t/2}\circ\mathcal{P}^{\mathrm{B}}_{\Delta t/2}, \tag{2.71}$$

where each semi group operator is labelled as,

$$\mathcal{P}^{\mathrm{X}}_{\Delta t} = e^{\Delta t\mathcal{L}_{\mathrm{X}}}. \tag{2.72}$$

Note that the flow associated with each operator in (2.70) can be solved analytically, and that a single timestep evaluation of (2.71) only requires a single gradient evaluation. This means that the method is simple to implement which together with the excellent statistical properties, shown in [47], have contributed to the rise in popularity of this method for MD. In fact, by combining (2.71), (2.68) and BCH it can be shown that any scheme following the pattern "XYZYX" gives, in general, a second order global weak error. From this we get,

$$\mathcal{P}^{\mathrm{B}}_{\Delta t/2}\circ\mathcal{P}^{\mathrm{A}}_{\Delta t/2}\circ\mathcal{P}^{\mathrm{O}}_{\Delta t}\circ\mathcal{P}^{\mathrm{A}}_{\Delta t/2}\circ\mathcal{P}^{\mathrm{B}}_{\Delta t/2} = e^{\Delta t\mathcal{L}+\Delta t^3\bar{\mathcal{L}}+\mathcal{O}(\Delta t^4)}, \tag{2.73}$$

where $\bar{\mathcal{L}}$ is an operator formed by commutators of $\mathcal{L}_{\mathrm{A}}$, $\mathcal{L}_{\mathrm{B}}$ and $\mathcal{L}_{\mathrm{O}}$. Expanding the right hand side of (2.73), we find the operators denoted as in (2.68) for this numerical method as,

$$\mathcal{A}_1 = \mathcal{L}, \qquad \mathcal{A}_2 = \mathcal{L}^2, \qquad \mathcal{A}_3 = \mathcal{L}^3 + \bar{\mathcal{L}}. \tag{2.74}$$

It is clear from this equation that the first permutation is introduced at order Δt^3 and we say that the BAOAB method has global weak second order. This method and the high-friction-limit method $\gamma\to\infty$ were rigorously analyzed in [47] and we refer to this work for more thorough evaluation than the one presented here.

In Figure 2.2 we present a trajectory for the uneven double well potential in two dimensions, using a friction of $\gamma = 1$ and target temperature $T = 1$. As this method is simple to implement and has excellent statistical properties, we have favoured this method and focus on building accelerated sampling schemes based on it. In particular, we derive a barostat based on this scheme which in a limit collapses to the original BAOAB scheme presented in this section.

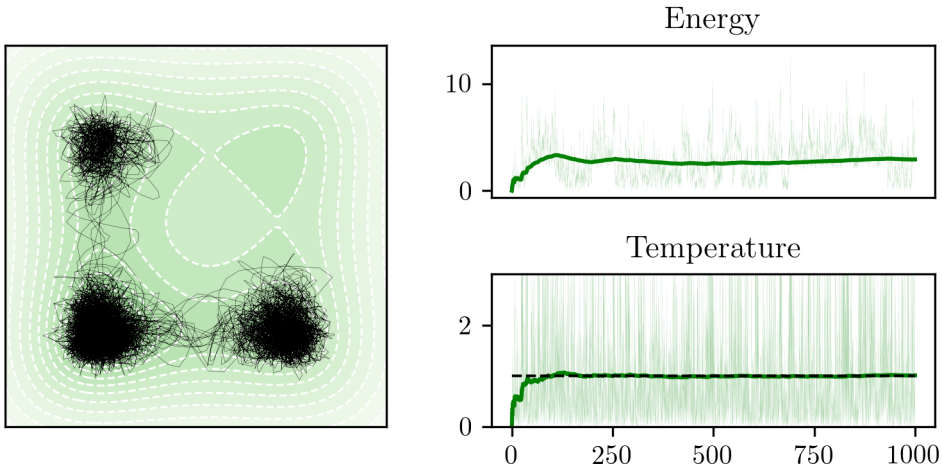

Figure 2.2: A trajectory is shown for the BAOAB scheme using a timestep of $h = 0.1$ and time length $t = 1000$. The long term average is indicated by the thick line and the instantaneous values are shown in the background. The expected temperature $T = 1$ in this case is shown as dashed black.

Chapter 3

An Overview of Accelerated Sampling Schemes

The field of accelerated sampling is vast, with a huge range of methods introduced over the last seven decades [50, 51, 52, 53, 54, 31]. In this section the intention is not to account for all these methods and strategies, but simply to summarize a small family of relevant schemes leading ultimately to the infinite switch schemes which will be the subject of study in Chapters 4 and 5.

In large-scale applications of sampling the potential energy landscape has a complex structure with many local minima, that trap systems locally and give rise to metastability. Traditional methods, such as the ones presented in the previous chapter, converge slowly for such problems and are unfortunately ill-suited for many real world applications. Currently, and for the foreseeable future, the calculation of high-dimensional integrals can only be performed efficiently via MCMC sampling methods and it is for this reason that the development of accelerated sapling methods is of great importance.

In Figure 3.1 the limitations of the Langevin method become apparent when applied to the Alanine-12 bio-molecule. The two states labelled A and B are separated by a large energetic barrier and although both states are prevalent at 300K. A researcher initialising her experiment near state A will come to the conclusion that the only configuration which exists at this temperature is A. If a second researcher conducts the same experiment but initalises his simulation near state B he will claim that the only stable state at this temperature is B.

Theoretically speaking, we know, as mathematicians, that given infinite computational resources both states will eventually reveal themselves to both researchers because they are using ergodic methods. In practice this is not seen because the convergence to the equilibrium average is too slow. The finite computational budget available to a researcher is too small to rely on ergodicity. For this reason one is forced to apply accelerated methods that converge faster to the invariant distribution.

Since accelerated methods are used for problems of complex energetic landscapes in high dimensions, to be effective, they must be able to generate reasonably accurate sampling results with limited computational expense. We recall that for a sequential sampling method such as a discretised stochastic differential equation the overall computational cost grows linearly with the number of required force evaluations per timestep, a property that typically rules out methods relying on higher order gradients, etc. In practice one finds that the best results are obtained by working near the stability threshold of numerical schemes. Some success has been obtained using methods that rely on extension of the phase-space. This might at first seem counter-intuitive since

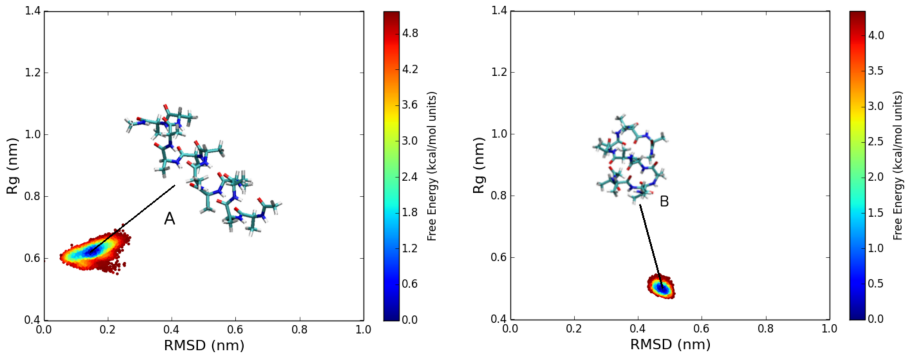

Figure 3.1: Reprinted with permission of the authors Bethune et al. [33]. (Left) Shows the free energy of the radius of gyration and root mean square distance, using Langevin dynamics at 300K, starting from the helical initial condition A. (Right) Shows the free energy in the same variables using Langevin dynamics at 300K, but instead starting from the initial configuration B.

it further increases the dimension of the problem, which is already initially assumed to be of high dimension.

The lack of exploration of the energetic surface can be viewed as the absence of exchange of information between different states on the landscape. Here, regions of high probability, state A and B from Figure 3.1, are separated by areas of low probability. By extending phase-space with a variable along which information can easily flow, regions of high probability can be connected through regions of low probability and transitions between metastable states are possible.

The simplest and most natural of these variables to use in molecular dynamics simulations is the temperature. At a high temperature the system can transition more easily between regions of metastability, as it possesses the required energy to overcome the energetic barrier. Methods based on this approach must define the flow of information between different temperature states such that the convergence to the equilibrium density is accelerated. Below we present a selection of methods that only differ in this respect.

3.1 Simulated Annealing

Simulated annealing was introduced by Kirkpatrick et al. in 1983 [55] and is not a sampling method but an optimization scheme designed to find the global minimum of a meta-stable function with a large number of parameters. Originally named after the *annealing* process used in metallurgy, the method has been used to solve many different problems (with the earliest application being to the travelling salesman problem [55, 56]). Annealing is the process of heating a metal and slowly cooling it to change the properties of the metal by re-crystallization [57].

In optimization the same process can be applied by introducing an effective temperature that together with some sampling scheme (originally Metropolis-Hastings) samples the instantaneous temperature by treating it as a temporary equilibrium. This should allow the model to effectively melt allowing the parameters to recrystallize – finding a new, ideally improved combinatorial solution.

As should be clear by now, large dimensional optimization and sampling of free energy landscapes generates multimodal landscapes. Assume that the parameter opti-

mization problem of interest is trapped in one such minimum and that our goal is to determine the global minimum of the system. With the fictitious temperature in place, the temperature is raised until the system effectively melts. Once this is achieved the cooling process begins.

The system will be cooled by using some predetermined cooling rate, which is a tuning parameter. This can be achieved in a number of ways, but it should allow the practitioner to control the cooling rate. Here, let the temperature at time fraction $t \in [0, 1]$ be the fraction of total simulation time. Define,

$$T(t, \lambda) = \frac{e^{-\lambda t} - e^{-\lambda}}{1 - e^{-\lambda}} T_{\max}. \tag{3.1}$$

Here, $T_{\max}$ is the highest temperature used and λ is the predetermined cooling rate. This function will cool the temperature from $T = T_{\max}$ at $t = 0$ to $T = 0$ at $t = 1$.

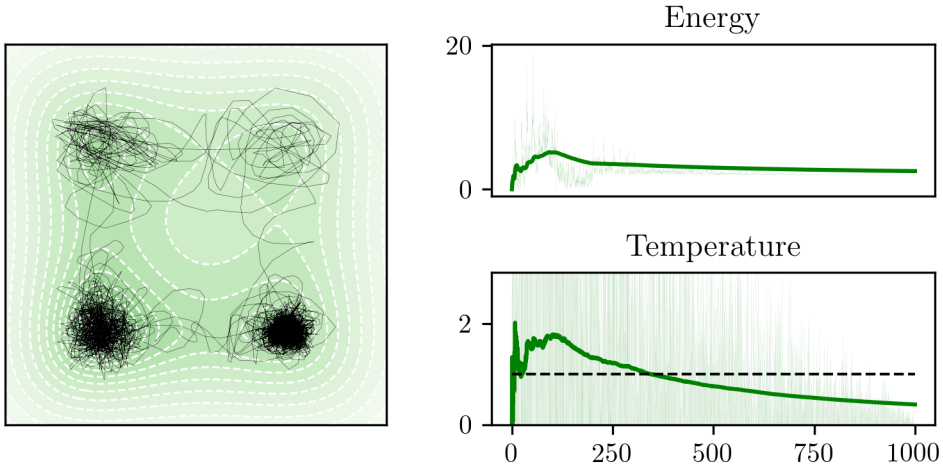

Figure 3.2: The figure shows the trajectory of a BAOAB integration with simulated annealing using a timestep of $h = 0.1$ and time interval length 1000. $T_{\max} = 2$ and $\lambda = 5.0$.

The simulated annealing method requires careful tuning of the cooling parameter to generate adequate performance [58], which may be difficult for a general problem. Letting the system cool too rapidly will impede the exploration and could generate a solution which gets trapped in a local minimum. Since the method is not designed to sample but to converge to the global minimum, one cannot actually claim to have found the global minimum unless the temperature is allowed to decay extremely slowly. In practice this method can therefore be inefficient and unreliable for large scale applications.

In the left panel of Figure 3.2 we display a trajectory for the 2D uneven double well potential. The high temperature has allowed the trajectory to explore all the four meta stable states, albeit briefly in the top right case. It also appears that the trajectory has become trapped in the lower right meta stable region, which is not the global minimum which is likely a result of cooling the system too quickly.

The cooling must be tuned such that the system is allowed to equilibrate at energy levels separating two meta stable states, such that the system is allowed time to escape into the lower of the two states. Continually cooling the system slowly around the these barriers should eventually lead the system to settle in the global minimum.

3.2 Simulated Tempering

In this section we discuss the simulated tempering method which was first introduced in 1992 by Marinari et.al [52]. This method draws inspiration from simulated annealing in the sense that it treats temperature as a dynamical variable. As was described in the previous section and was clear from the simulation, simulated annealing heavily depends on the cooling rate. It also relies on an ad hoc cooling schedule and in practice it is unlikely that a realistic simulation on a finite resource computer would find the global minimum. To remedy this issue, often, many different simulations are conducted which are initialized at different initial points. Even though increasing the number of trajectories aids in the exploration of the energy landscape, there is no guarantee that any individual trajectory would find the global minimum [52].

Simulated tempering extends this approach by allowing any trajectory to either increase or decrease its temperature. The heuristic gain from allowing the trajectory to jump in temperature can be understood if we consider a the exploration of a rugged potential $U(\boldsymbol{q})$. This potential will have a large number of local minimums, that at any point can constrain the trajectory in its local neighbourhood such that the trajectory never discovers the global minimum. By allowing a trajectory to increase its temperature, it can escape the local minimum and continue exploring the potential landscape $U(\boldsymbol{q})$. The goal in simulated tempering is in fact different than that in simulated annealing: the purpose has shifted to canonical sampling from global optimization.

Simultaneously, metastable states are also regions of high-probability. Using artifically elevated temperatures would "flatten" the landscape so that the metastable states are under-resolved. By simultaneously allowing for temperature jumps in both directions, the high temperatures will aid the exploration whilst the low temperatures will allow for the basins to be resolved. For simple systems involving only energetic barries, one can expect artificial heating to improve the exploration rate. On the other hand, increasing the temperature to overcome an entropic barrier does not help the simulation to traverse it. The simulated tempering method is therefore only expected to produce good results for systems with energetic and not entropic barriers.

Since the goal of the simulated tempering method is to accelerate the sampling of the canonical measure, the dynamics of temperatures must be conducted in such a way that the simulated tempering trajectory can be unbiased [59]. By treating the temperature as a random walk the extended conserved measure is known and a specific implementation of the method can be unbiased to find the canonical measure at any temperature.

Let us now introduce the method. Consider a monotone increasing sequence of K temperatures between $[T_{\min}, T_{\max}]$,

$$T_1 = T_{\min} < T_2 < \ldots < T_K = T_{\max}. \tag{3.2}$$

This sequence is often referred to as a temperature "ladder". Define reciprocal temperatures $\beta_i = 1/k_B T_i$ for all $i \leq K$ and specify for each such reciprocal temperature a weight $\omega(\beta_i) > 0$. The proposal positions and potential momenta sampled from the canonical distribution $\rho_{\beta_i}(\boldsymbol{q}, \boldsymbol{p})$, is undertaken with the practitioner's favourite method. The full distribution that such a process must sample is given by,

$$\pi(\boldsymbol{q}, \boldsymbol{p}, \beta) \propto \pi(\boldsymbol{q}, \beta) \propto \sum_{i=1}^{K} \omega(\beta_i) e^{-\beta_i U(\boldsymbol{q})} \, \mathrm{d}\boldsymbol{q} \delta_{\beta_i}(\mathrm{d}\beta). \tag{3.3}$$

The instantaneous temperature for any trajectory sampling (3.3) is selected according to the following Metropolis-Hastings algorithm:

1. Integrate the dynamics $(\boldsymbol{q}, \boldsymbol{p})$ at temperature β_i on the time interval $[0, \tau]$ with timestep $\Delta t = \tau/n$ to obtain $(\boldsymbol{q}_n, \boldsymbol{p}_n)$. Set $(\boldsymbol{q}, \boldsymbol{p}) = (\boldsymbol{q}_n, \boldsymbol{p}_n)$

2. Generate the proposal temperature $\tilde{\beta}_i$ with probability $1/K$ and accept this new proposal i.e. set $\beta_{i+1} = \tilde{\beta}_i$ with probability,

$$\min\left(1, \frac{\omega(\tilde{\beta}_i)}{\omega(\beta_i)} \exp\left[-\left(\tilde{\beta}_i - \beta_i\right) U\left(\boldsymbol{q}\right)\right]\right), \tag{3.4}$$

   otherwise let $\beta_{i+1} = \beta_i$.

It is clear from this outline that for a temperature switch to occur the probability of switching from one temperature to the next cannot be negligible [60]. For practical implementations and to increase the effectiveness of the algorithm, proposals are only generated as a bounded random walk. This implies that in practice, often temperature switches are only proposed between neighbouring temperatures with a probability $1/2$ to either jump up or down (or remaining at the current temperature or increasing if at $T_{\min}$ or decreasing for $T_{\max}$).

The number of temperatures and their distribution cannot be known a priori and thus becomes a tuning parameter. A large temperature domain will allow for the use of a higher temperature which aids the exploration. However, a larger domain requires a larger total number of temperatures to cover it with a significant overlap to allow for movement in temperature up and down the ladder.

Let $\varphi\left(\boldsymbol{q}\right)$ be some observable. To collect statistics for the empirical estimate of $\mathbb{E}_{\beta_i}\left[\varphi\right]$, for some temperature T_i in the ladder, it is common to independently record statistics of $\mathbb{E}_{\beta_i}\left[\varphi\right]$ at each of the K steps of the ladder [61]. This implies that there is a trade-off between the exploration (proposing a temperature switch) and collecting statistics for the current temperature. This manifests itself in the tuning of the time interval τ which if sufficiently short aids exploration and if long aids the collection of statistics [62]. In the work undertaken in [31] and outlined in Chapter 4 we show how the optimal limit is to take $\tau \to 0$: i.e.. to use infinitely fast switching between temperatures.

The performance of any given number and distribution of temperatures is difficult to predict for any given problem. Likewise, the knowledge of the types of barriers in the potential landscape ultimately also affects the decisions of the method parameters. Tuning these variables to achieve a satisfactory efficiency is a difficult task which involves some intuition about the problem at hand [63].

One of the more involved decisions is to determine the weight factor $\omega(\beta_i)$ associated with each temperature [61]. This choice is often motivated by the probability of observing any temperature β_i. Assume that there is a constant reference temperature $\beta \in [\beta_{\min}, \beta_{\max}]$ and that at after a time length τ the force is rescaled by a factor β_i/β. Then a simulated tempering trajectory explores the measure (3.3). Let $Z_{\boldsymbol{q}}\left(\beta_i\right)$ be the configurational part of the partition function of (3.3). A simple calculation reveals that the marginal distribution in temperature of (3.3) for any $\beta_i \in [\beta_{\min}, \beta_{\max}]$ is,

$$\rho(\beta_i) = \frac{\omega(\beta_i) Z_{\boldsymbol{q}}\left(\beta_i\right)}{\sum_{i=1}^{M} \omega(\beta_i) Z_{\boldsymbol{q}}\left(\beta_i\right)}. \tag{3.5}$$

It is consequently clear that the marginal distribution in temperature is uniform if $\omega(\beta_i) \propto Z_{\boldsymbol{q}}^{-1}(\beta_i)$. The uniform feature of this marginal distribution is desirable as it implies that all the temperatures in the range are sampled with equal probability. This is beneficial for the collection of statistics as an equal number of points is sampled from any temperature, and that the computational effort is expended equally across all temperatures without oversampling either the low or high limit. An alternative to this choice is described in Lelièvre and Stoltz [7, s. 3.1], where the weight $\omega(\beta_i)$ is chosen as the inverse of the variance of the integral in time at each β_i such that the statistical errors are minimised.

Naturally, this complicates the implementation and deployment of the algorithm as the partition function is not known in general and is difficult to calculate. Indeed if it was known, the exercise of sampling becomes redundant as the canonical measure is known up to this constant and as such any observable could be calculated.

On the other hand, it is of course possible to make a different choice of the weight fraction in which case one could chose an a priori known function. This choice could be used to bias the sampling towards a desired temperature regime. Depending on the particular application this could potentially be based on intimate knowledge of the system of study. Practically, such information is hard to derive prior to an extensive computational study and practitioners typically resort to using the partition function as the weight fraction.

The required knowledge of the partition function is the reason why the method has lost popularity [64] to "parallel tempering" (discussed in Section 3.3) which does not depend upon this information. Commonly, an adaptive learning procedure is implemented alongside the dynamical temperature process. This results in adaptively computing free-energy differences as is also done in Wang-Landau [65] or ABF [66]. We have also introduced our own learning procedure [31] which is discussed in Section 4.1.6 and this could be coupled with the vanilla simulated tempering algorithm of this section.

Below we present an on-the-fly version of the weight determination (adaptive computing of free-energy differences) introduced by Zhang et. al [67] which is a generalisation of the recurrence scheme first proposed by Park and Pande [61]. This method considers only switches between two consecutive temperature states and tries to approximate a weight which typically satisfies detailed balance of a temperature exchange in simulated tempering. Introduce the two neighbour states β_1 and β_2 and recall that a proposal to switch from $\beta_1 \to \beta_2$ is accepted with probability,

$$\min\left(1, \exp\left[-\left(\beta_2 - \beta_1\right) U\left(\boldsymbol{q}\right) + \log \omega(\beta_2) - \log \omega(\beta_1)\right]\right). \tag{3.6}$$

Note that this ratio only relies on knowing the difference (or fraction) between the two weights and not the weights themselves. Now, we are seeking an approximation of the weight which is equal to the reciprocal partition function,

$$\omega(\beta) = Z^{-1}(\beta). \tag{3.7}$$

Introduce the averaged potential energy at any temperature β in the range $[\beta_{\min}, \beta_{\max}]$ and define this as,

$$\bar{U}(\beta) = \frac{\int_\Omega U\left(\boldsymbol{q}\right) e^{-\beta U(\boldsymbol{q})} \mathrm{d}\boldsymbol{q}}{\int_\Omega e^{-\beta U(\boldsymbol{q})} \mathrm{d}\boldsymbol{q}} = -\frac{Z'(\beta)}{Z(\beta)}. \tag{3.8}$$

Differentiating (3.7) and combining this with (3.8) one can solve the resulting equation

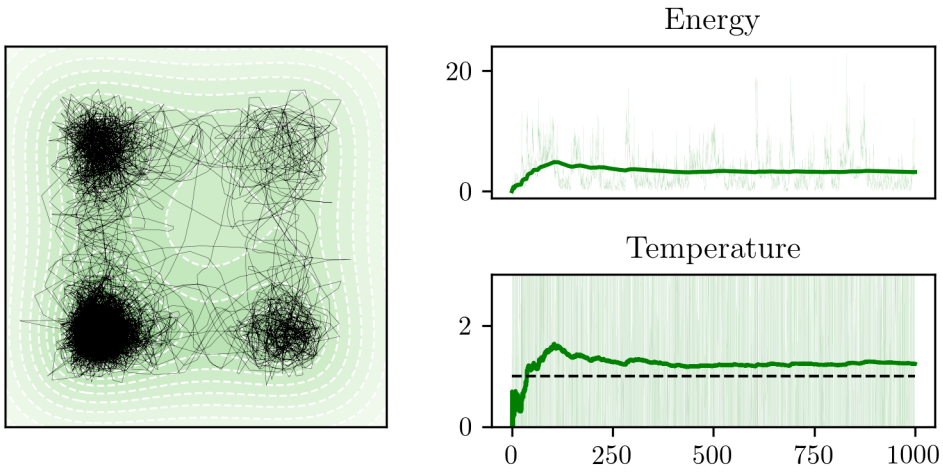

Figure 3.3: The figure shows the trajectory of a BAOAB with a simulated tempering trajectory using a timestep of $h = 0.1$ and time length $t = 1000$. $T_{\min} = 0.5$ and $T_{\max} = 2$ with 10 temperatures distributed linearly in β with a proposed temperature switch at every step.

for $\omega(\beta)$ to obtain,

$$\log \omega(\beta) = \log \omega(\beta_{\min}) + \int_{\beta_{\min}}^{\beta} \bar{U}\left(\beta'\right) \mathrm{d}\beta'. \tag{3.9}$$

Consider again the switch from $\beta_1 \to \beta_2$. In this case, one can estimate the integral in (3.9) as a simple sum to obtain,

$$\log \omega(\beta_2) = \log \omega(\beta_1) + (\beta_2 - \beta_1)\left(\bar{U}(\beta_2) - \bar{U}(\beta_1)\right)/2. \tag{3.10}$$

Determining $\bar{U}$ empirically at each temperature one can bootstrap to approximate relative weights between any two neighbouring states. This has previously been discussed [61] and a very short argument is provided in that work for why this satisfies detailed balance. The fully adaptive scheme, complete with a dynamical bootstrapping strategy, can be found in [67] and is shown for a short trajectory in Figure 3.3.

The simulated tempering method with adaptive weight learning [67] is shown in Figure 3.3 combined with a BAOAB trajectory. The temperature and energy shown in the right panels are not physical and show the values obtained by the trajectory sampling the full measure (3.3) and is therefore not reweighted. Values for any observable can be obtained by discretization in temperature as explained above–this is trivial to implement and can also be done in "post production".

This run should be compared to the standard Langevin implementation shown in Figure 2.2 and the comparison illustrates the improved exploration gained by introducing the uniform random walk in temperature. For example, it is clear from Figure 3.3 that the trajectory has explored the shallow metastable state in the top right hand corner. Although a simple toy problem, the difference between vanilla Langevin and simulated tempering Langevin are clearly illustrated for this system where the advantage of simulated tempering is apparent.

3.3 Parallel Tempering / Replica Exchange

The *replica exchange method* (REM), also known as parallel tempering, is an accelerated sampling scheme that scales well vertically i.e spread across many cores. First introduced by Swendsen and Wang in 1986 [68], the method was used for many years in conjunction with MCMC sampling until 1999 when Sugita and Okamoto [69] first used it together with MD.

The method can in fact be considered to be a relative of simulated tempering since it also uses a ladder of temperatures to accelerate the sampling. The differences between the methods lie in the number of replica systems used and how information is exchanged between the temperature levels. In ST there is one replica which occupies a single temperature and the system experiences a change in this temperature as a result of a random walk in temperature space. This is in contrast to REM, which instead holds an equal number of temperatures and replicas, such that there is at all times a one to one correspondence between them. A temperature is then exchanged between any two replicas, permuting the temperature index each replica is assigned over time. This might appear as a semantic difference but, as we will see below, generates a significant difference to ST in how the acceptance ratio is calculated and as a result, no prior knowledge of the partition function is required.

This difference in the calculation of the acceptance probability between ST and REM is the most likely explanation of why REM has been the prefered method of choice in literature over ST [70]. The method also parallelizes trivially, which results in an extremely good vertical scaling on multicore computing systems, whilst maintaining a simple implementation structure [71]. Together with the rise of such computer systems, REM appears to have aged well despite being almost 34 years old. As a natural consequence, it has evolved and several improvements on the original method have been made [64, 70, 72].

Let us first discuss the original algorithm and its implementation. The heuristic reasoning behind the algorithm is the same as that of ST: high temperature states explore the state space and low temperature states provide the necessary resolution of high probability regions. Consider to this end a monotone increasing sequence of K temperatures in $[T_{\min}, T_{\max}]$,

$$T_1 = T_{\min} < T_2 < \ldots < T_K = T_{\max}. \tag{3.11}$$

Let the reciprocal temperature be $\beta_k = 1/k_B T_k$. Introduce K replicas of the system $(\boldsymbol{q}, \boldsymbol{p})$ and assign each such replica a temperature. Denote the replica at temperature $0 \leq k \leq K$ as $(\boldsymbol{q}_k, \boldsymbol{p}_k)$. Define the canonical measure for replica k at temperature with index j as,

$$\mu_{\beta_j}\left(\mathrm{d}\boldsymbol{q}_k\, \mathrm{d}\boldsymbol{p}_k\right) = Z_{\beta_j}^{-1} \exp\left[-\beta_j H\left(\boldsymbol{q}_k, \boldsymbol{p}_k\right)\right] \mathrm{d}\boldsymbol{q}_k\, \mathrm{d}\boldsymbol{p}_k, \tag{3.12}$$

with density,

$$\rho_j\left(\boldsymbol{q}_k, \boldsymbol{p}_k\right) = Z_{\beta_j}^{-1} \exp\left[-\beta_j H\left(\boldsymbol{q}_k, \boldsymbol{p}_k\right)\right]. \tag{3.13}$$

This holds for all replicas as only one replica can occupy any temperature at a given point in time. It will be useful to think of a state (replica to temperature maps) as a permutation of the K indices. This implies that, one can think of the normal phase space to be extended with index space. Let σ be the index permutation of the K indices that determines the current configuration state for the K replicas. These maps define the temperatures which each replica is assigned to. At any point in time, the system of replicas will be in such a configuration. Consider the marginal of the extended

distribution,

$$\rho\left(\boldsymbol{q},\boldsymbol{p},\sigma\right)=\frac{1}{K!}\prod_{k=0}^{K}\rho_{\sigma(k)}\left(\boldsymbol{q}_k,\boldsymbol{p}_k\right) \tag{3.14}$$

such that the density of any extended distribution or index permutation σ is,

$$\rho\left(\sigma|\boldsymbol{q},\boldsymbol{p}\right)=\frac{\rho\left(\boldsymbol{q},\boldsymbol{p},\sigma\right)}{\sum_{\sigma'}\rho\left(\boldsymbol{q},\boldsymbol{p},\sigma'\right)}. \tag{3.15}$$

Working in this setting, phase-space plus index space, we now define the REM algorithm for $(\boldsymbol{q},\boldsymbol{p})$ fixed. Introduce the transition kernel $T_{\sigma,\sigma'}\left(\boldsymbol{q},\boldsymbol{p}\right)$, defining the transition from index map $\sigma\to\sigma'$. Assume that this kernel satisfies detailed balance with respect to $\rho\left(\sigma|\boldsymbol{q},\boldsymbol{p}\right)$, meaning

$$\rho\left(\sigma|\boldsymbol{q},\boldsymbol{p}\right)T_{\sigma,\sigma'}\left(\boldsymbol{q},\boldsymbol{p}\right)=\rho\left(\sigma'|\boldsymbol{q},\boldsymbol{p}\right)T_{\sigma',\sigma}\left(\boldsymbol{q},\boldsymbol{p}\right). \tag{3.16}$$

In practice the simplest such choice is a symmetric kernel with uniform sampling over the index space, so that the probability of picking any configuration is equivalent. Similarly, to make the algorithm practically attainable with adequate performance, one only considers exchanging one pair–often also at neighbouring temperatures. Picking any such pair uniformly is allowed within the framework given above and only improves the performance of the algorithm without affecting the analysis.

Let σ_i be the i^{th} index permutation iteration and define the evolution of the permutation according to the following algorithm:

1. Integrate the dynamics for all the replicas at their assigned temperatures, i.e. replica $(\boldsymbol{q}_k,\boldsymbol{p}_k)$ at temperature $\beta_{\sigma_i(k)}$ on the time interval $[0,\tau]$ with timestep Δt.

2. Generate a proposal permutation $\tilde{\sigma}$ with probability $1/(K-1)$ of exchanging any neighbouring pair. Accept this new proposal i.e. set $\sigma_{i+1}=\tilde{\sigma}$ with probability,

$$\min\left(1,\exp\left[\left(\beta_{\sigma(k)}-\beta_{\tilde{\sigma}(k)}\right)\left(H\left(\boldsymbol{q}_{\sigma_i(k)},\boldsymbol{p}_{\sigma_i(k)}\right)-H\left(\boldsymbol{q}_{\tilde{\sigma}(k)},\boldsymbol{p}_{\tilde{\sigma}(k)}\right)\right)\right]\right), \tag{3.17}$$

   otherwise let $\sigma_{i+1}=\sigma_i$

Note, the difference in this acceptance probability (3.17) compared to the one used in simulated tempering (3.4) and the lack of required knowledge of any weight function in the first. As was discussed in the previous section, this weight function should be chosen as the inverse of the partition function for uniform sampling in temperature. This type of knowledge is not necessary to obtain in order to perform simulations with REM. This is a consequence of the definition of (3.17), which is derived by enforcing detailed balance (3.16) with invariant measure (3.15). Under the assumption that the transition kernel is uniform the acceptance probability is,

$$\min\left(1,\frac{\rho\left(\boldsymbol{q},\boldsymbol{p},\sigma\right)}{\rho\left(\boldsymbol{q},\boldsymbol{p},\tilde{\sigma}\right)}\right), \tag{3.18}$$

for proposal index map $\tilde{\sigma}$. This makes it immediately clear why no prior knowledge of any of the partition functions are required.

Up to this point we have not discussed the choice of temperatures, which influences the overlap between consecutive temperature and therefore the probability of exchanging two neighbouring replicas. This implies that the overall effectiveness of the algorithm is directly dependent on how the temperatures are distributed. Equally, a large

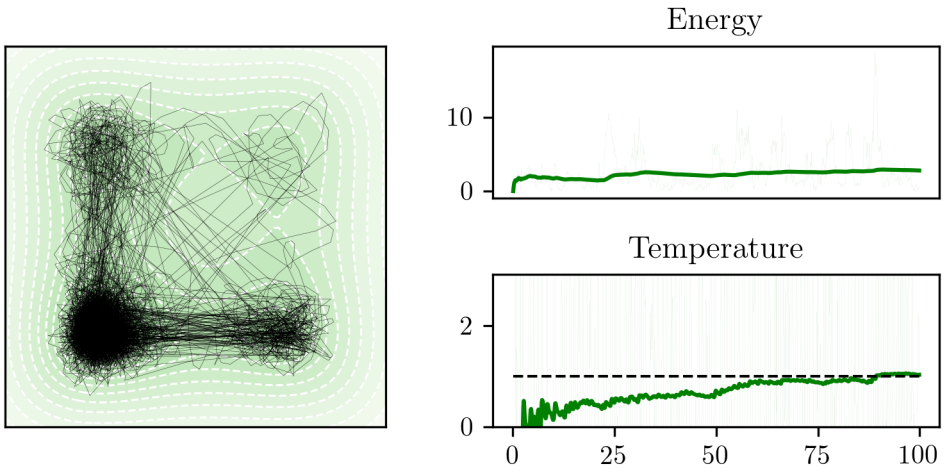

Figure 3.4: Shows the trajectories of 10 replicas of time length $t = 100$ using a timestep of $h = 0.1$ with a temperature proposal exchange on every 10 steps with momentum reweighting. Here $T_{\text{min}} = 0.5$ and $T_{\text{max}} = 2$ with one replica at each temperature at any point in time with the temperatures distributed linearly in β. The plots in the right panels show the averages of replica 0.

number of replicas and temperatures increases the computational cost but enhances the exploratory properties of the algorithm. Depending on the desired temperature range which one wishes to cover, a larger number of temperatures could even be necessary to achieve good performance [70].

Indeed, the use of too many replicas will affect the acceptance probability (3.17) as the probability of a switch is directly dependent on the change in temperature $\Delta\beta$. This implies that if this quantity becomes too small, the probability of a successful switch will diminish with the differences in the temperature ladder. Since REM is a very popular accelerated sampling scheme, the problem of optimal distribution of temperatures has been studied before [73, 74]. The consensus in the literature has been to use a geometric distribution of reciprocal temperature β.

In Figure 3.4 we show a simple implementation of the replica exchange method with a total of 10 replicas allocated to a respective temperature, linear distributed in β. We reiterate that these pictures are only shown as illustrative examples of the different methods, in the absence of any significant tuning. We also stress that it is not necessary to exchange the coordinates, but one can simply only exchange the temperature between the replicas. However, in Figure 3.4 the trajectories have been exchanged to emphasize the algorithm's properties which are evident from the straight lines connecting the metastable states, all representing an exchange of coordinates.

To keep the computational cost comparable to the previously presented methods, each trajectory is 10 times shorter than in the previous sections, such that the number of force evaluations per timestep is kept constant. The values for the (arbitrarily chosen) first replica are shown in the right panels in Figure 3.4. The replica exchange method was combined with BAOAB but could have been combined with any other canonical sampling scheme.

Similar to simulated tempering, one must re-weight the trajectories to obtain statistics at the different temperature levels. This is simply done by recording the relevant statistic into the correct histogram. Consider the observable $\varphi(\boldsymbol{x})$ where $\boldsymbol{x}$ is a vector

in the same space as any replica $\boldsymbol{q}_k$ and similarly for the conjugate momentum $\boldsymbol{p}_x$. The simple histogram reweighting of the expectation of this observable at temperature j is

$$\begin{aligned}\mathbb{E}_j\left[\varphi\left(\boldsymbol{x}\right)\right] &= \int \varphi\left(\boldsymbol{x}\right)\rho_j\left(\boldsymbol{x},\boldsymbol{p}_x\right)\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{p}_x, \\ &= \int \sum_{\sigma}\sum_{k}^{K}\varphi(\boldsymbol{q}_k)\mathbb{1}_{j=\sigma(k)}\rho\left(\boldsymbol{q},\boldsymbol{p},\sigma\right)\mathrm{d}\boldsymbol{q}\mathrm{d}\boldsymbol{p}.\end{aligned} \tag{3.19}$$

This quantity is straightforward to calculate.

Note that the switching period τ seems at first glance to be somewhat of an elusive parameter to determine the optimal value for. Empirical studies have shown that letting $\tau \to 0$ i.e. sending the switching frequency to infinity, improves the sampling performance [75]. Evaluating the dynamics with an increasing number of acceptance-rejection steps could mean that a significant portion of the computational budget is wasted on this task, with little overall gain in sampling performance.

3.3.1 Accelerating Accelerated Sampling

In 2012 Dupuis et al. [32] showed that it is possible to numerically implement parallel tempering with an infinitely fast swapping rate. The resulting equations of motion are actually swapping free and consist of modifying the diffusion coefficient which results in SDEs with multiplicative noise terms. Importantly, the authors of this paper show, by using a large deviation argument [76, 77, 78], that the new set of SDEs has a faster convergence rate to the invariant distribution, than the original equations. Practically this method is also an improvement since any potentially wasted computational effort on calculating acceptance rates is avoided.

Building on the analysis of this work, Lu and Vanden-Eijnden [70] showed that it is possible to avoid a geometric noise component. A complete analysis of this work was recently published in [79] on which we base the arguments outlined below.

The argument presented in [32] is simple and consists of looking at the probability that the empirical measure at time T is different from' the invariant target measure μ, which is accessible through the following large deviation argument. Let X_1 and X_2 denote two replicas satisfying independent overdamped Langevin processes,

$$\begin{aligned}\mathrm{d}X_1 &= -\nabla U(X_1)\,\mathrm{d}t + \sqrt{2\beta_1}\,\mathrm{d}W_1, \\ \mathrm{d}X_2 &= -\nabla U(X_2)\,\mathrm{d}t + \sqrt{2\beta_2}\,\mathrm{d}W_2.\end{aligned} \tag{3.20}$$

The equations naturally suggest a joint probability density of the form,

$$\rho(x_1,x_2) \propto e^{-\beta_1 U(x_1)}e^{-\beta_2 U(x_2)}. \tag{3.21}$$

This implies that we can consider two temperature index permutations: σ,σ' where $\sigma(1)=1$, $\sigma(2)=2$ and $\sigma'(1)=2$, $\sigma'(2)=1$. Associated with these two configurations are the conditional densities for each respective permutation: $\varrho(x_1,x_2|\sigma)=\rho(x_1,x_2)$ and $\varrho(x_1,x_2|\sigma')=\rho(x_2,x_1)$. If we consider the probability of finding the system in either of the two configurations, it is natural to define the density,

$$\varrho(\sigma|x_1,x_2) = \frac{\varrho(x_1,x_2|\sigma)}{\varrho(x_1,x_2|\sigma)+\varrho(x_1,x_2|\sigma')}. \tag{3.22}$$

Let $\nu \in (0,\infty)$ denote the swapping intensity of the Markov jump process that ex-

changes the temperatures between the replicas and has the transition rate $T_{\sigma,\sigma'}(x_1, x_2) \geq 0$, satisfying the detailed balance condition w.r.t. (3.22),

$$T_{\sigma,\sigma'}(x_1, x_2)\varrho(\sigma|x_1, x_2) = T_{\sigma',\sigma}(x_1, x_2)\varrho(\sigma'|x_1, x_2). \tag{3.23}$$

The generator of this process can be written down for an observable $\varphi(x_1, x_2, \sigma)$, where we make the dependence on the current configuration explicit through σ. The generator of the combined process $\mathcal{L}_\nu$ is written as,

$$\mathcal{L}_\nu\varphi(x_1, x_2, \sigma) = \mathcal{L}^\sigma\varphi(x_1, x_2, \sigma) - \nu T_{\sigma,\sigma'}(x_1, x_2)\left(\varphi(x_1, x_2, \sigma) - \varphi(x_1, x_2, \sigma')\right), \tag{3.24}$$

with,

$$\mathcal{L}^\sigma\varphi = \nabla_{x_1}U \cdot \nabla_{x_1}\varphi + \beta_{\sigma(1)}\Delta_{x_1}\varphi + \nabla_{x_2}U \cdot \nabla_{x_2}\varphi + \beta_{\sigma(2)}\Delta_{x_2}\varphi. \tag{3.25}$$

Equation (3.24) makes the dependence on the switching frequency ν clear. The first term in (3.24) is the generator of the overdamped processes where the diffusion is determined by the permutation, σ.

The jump process (3.24) satisfies the detailed balance condition (3.23), and $\mathcal{L}^\sigma$ is the generator of the over damped Langevin processes. From this it can be shown that, with suitable conditions on the potential U satisfied, the invariant distribution of the process is $\rho(x_1, x_2|\sigma)$ such that,

$$\int_{\mathbb{R}\times\mathbb{R}} \mathcal{L}_\nu\varphi(x_1, x_2, \sigma)\varrho(x_1, x_2|\sigma)\,\mathrm{d}x_1\,\mathrm{d}x_2 = 0. \tag{3.26}$$

From this and the ergodic theorem we consider the empirical measure defined as,

$$\mu_T^{\nu,\sigma} = \frac{1}{T}\int_0^T \delta_{(X_1^{\nu,\sigma}(t),\ X_2^{\nu,\sigma}(t))}\,\mathrm{d}t, \tag{3.27}$$

where $X_1^{\nu,\sigma}(t)$ and $X_2^{\nu,\sigma}(t)$ denote two process with jump rate ν in configuration σ. It was first noted by Dupuis et al. [32] that the rate of convergence of this empirical measure to the invariant measure $\varrho(x_1, x_2|\sigma)$, can be characterised by using a large deviation principle. This principle is used to show that the difference between two probability measures is asymptotically small in the large time limit and is determined by the rate functional, often labelled $I(\mu)$. A larger rate functional indicates a more rapid convergence. To be more precise,

Definition 9 (Large Deviation Principle)**.** *Let $\Gamma \subseteq P\left(\mathbb{R}^d\right)$ where $P\left(\mathbb{R}^d\right)$ is the space of all probability measures on $\mathbb{R}^d$. A sequence of random probability measures $\{\gamma_T\}_{T\geq 0} \in \Gamma$ is said to satisfy a large deviation principle (LDP) with rate function I if,*

$$\lim_{T\to\infty}\frac{1}{T}\log\mathbb{P}\left\{\gamma_T \in \Gamma\right\} = -\inf_{\mu\in\Gamma} I(\mu). \tag{3.28}$$

Because of the additive structure in (3.24) it can be shown that the empirical measure (3.27) satisfies a large deviation principle with respect to $\varrho(x_1, x_2|\sigma)$, also of an additive form. In fact, this rate is given by,

$$I_\nu\left(\varrho\right) = J_0\left(\varrho\right) + \nu J_1\left(\varrho\right). \tag{3.29}$$

Here, J_0 determines the rate when no swap occurs whilst J_1 determines the jump rate at which point the replicas are exchanged. By observing that $I\left(\varrho\right)$ is a monotone increasing function in ν, it is possible to show that $I^\infty\left(\varrho\right) = \lim_{\nu\to\infty} I_\nu\left(\varrho\right) = \infty$ in

the limit as $\nu \to \infty$. This indicates that the convergence is improved as the switching frequency is increased and if it is possible to write down equations for a process with rate functional $I^{\infty}(\varrho)$ this would be optimal.

The outline presented in this Chapter forms the base for the work done in the next, were we investigate the infinite switch limit of the simulated tempering method – showing that the infinite switch limit is also optimal in this case.

Chapter 4

Infinite Switch Simulated Tempering with Adaptive Weight Learning

As we have seen in the last chapter, the development of algorithms that accelerate sampling of high-dimensional probability distributions is of great importance to a wide variety of computational fields. Many standard sampling methods in common use, such as those outlined previously in this thesis, fail to achieve adequate performance when the potential energy contains complex features such as multimodality. In practice this inadequacy means that practitioners may attain poor coverage of the physically relevant states $(\boldsymbol{q}, \boldsymbol{p})$. Missed states can ultimately result in inferior understanding of the macroscopic properties of the system. As an illustration, this is detrimental in drug design as potentially beneficial folding states of certain proteins remain undisclosed whilst a few states near the initial condition are over-sampled.

There are a wealth of methods that address this problem by improving the exploration of high-dimensional multi-modal probability densities. The method presented in this chapter was primarily designed to overcome energetic barriers that impede the exploration of the probability density. From now on, we will use *improved sampling* interchangeably with 'accelerating the exploration' of said densities. Because of the often high dimensionality of the systems studied, one of our main underlying design considerations must be to accelerate the sampling with minimal or no increase in computational cost.

Methods that improve sampling performance by extending the phase-space by treating temperature as a dynamical variable were introduced in the previous chapter: simulated annealing, replica exchange, and simulated tempering. Here, we will investigate the theoretical foundations of the simulated tempering (ST) method. As we saw earlier, the practical implementation of ST involves making a number of choices for different sampling parameters. These choices are highlighted in the following three tasks:

1. *Pick the number of temperatures M and determine their spatial distribution between some limit temperatures: $T_{\min}$ and $T_{\max}$.*

   There are a number of underlying issues to address in this case. The more temperatures that are used, the longer the sampling trajectory has to be to collect a sufficient number of samples at each temperature. Secondly, if the temperatures are distributed such that the probability of accepting a switch is vanishingly small, the sampling will not benefit from all the temperatures, and some temperatures

will likely never be visited. This will result in a negligible acceleration of the sampling given a fixed computational budget.

2. *Choose a sensible switching time τ.*

   In making this choice, a practitioner has to consider the number of temperatures and the desired sampling at each temperature. In addition, this does not necessarily have to be a constant but could also be chosen at random. However, in general experiments it has been noted that as $\tau \to 0$ the sampling performance is improved [62].

3. *Make a justified choice for the reciprocal temperature weight, $\omega(\beta)$.*

   This is the most problematic choice. On the other hand it is trivial to show that if $\omega(\beta) \propto Z_{\boldsymbol{q}}^{-1}(\beta)$ i.e. the weight is proportional to the configurational part of the partition function, the distribution is uniform in temperature and the probability of choosing any temperature is $1/M$. This is deemed desirable as no temperature level will be oversampled or favoured, which otherwise could impede the number of samples collected at each temperature.

In an effective implementation of the ST method we must make choices for all the parameters mentioned above. In particular it should be noted that Task 3 is the most problematic from an implementation perspective, as it requires knowledge of the configurational partition function for uniform sampling in temperature. This is the primary reason for the limited adoption of the method in practice.

In the work presented in this section we will aim to address the issues Tasks 1–3 by adapting a large deviation principle proposed by Dupuis et al in [80, 32] for replica exchange molecular dynamics. Simultaneously, we will work with a continuous range of reciprocal temperatures $[\beta_{\min}, \beta_{\max}]$. We will show using the aforementioned structure that the infinite switch limit of simulated tempering is equivalent to a rejection free dynamics sampling a Boltzmann distribution with the averaged potential

$$\bar{U}(\boldsymbol{q}) = -\beta^{-1} \log \int_{\beta_{\min}}^{\beta_{\max}} \omega(\beta_c) \exp\left[-\beta_c U(\boldsymbol{q})\right] \mathrm{d}\beta_c. \tag{4.1}$$

The reference β is the physical reciprocal temperature at which $\bar{U}(\boldsymbol{q})$ is sampled and β_c denotes the continuous reciprocal temperature that is integrated over $[\beta_{\min}, \beta_{\max}]$.

The use of a continuous temperature range for tempering was first introduced by Gobbo and Leimkuhler [81]. Their method differs form the one presented here in that they temper on an auxiliary parameter ξ, where the reciprocal temperature is given as a function of ξ, i.e.. $\beta_c(\xi)$, instead of direct tempering on β_c as done here. We also mention the similar integrate-over-temperature-method developed by Gao [82], which uses a set of discrete temperatures. In contrast we show that it is better to use a continuous range of β_c.

The foundation of the ISST method is presented in the next section and we derive the simplified equations for the physical variables $(\boldsymbol{q}, \boldsymbol{p})$, which eliminates the need to simulate temperature. Additionally, the choice of weight $\omega(\beta_c)$ is discussed.

4.1 Foundations

4.1.1 A Continuous Formulation of Simulated Tempering

Consider a continuously varying reciprocal temperature β_c in the interval $[\beta_{\min}, \beta_{\max}]$. This aids the analysis but does not affect the result, and all conclusions still hold starting from a discrete set $\{\beta_i\}_{0\leq i\leq M} \subset [\beta_{\min}, \beta_{\max}]$. Note also that the fixed physical temperature β should not be confused with the continuous β_c. This is made clear in the extended Gibbs distribution,

$$\rho_\beta\left(\boldsymbol{q},\boldsymbol{p},\beta_c\right) = Z^{-1}(\beta)\omega(\beta_c)\exp\left[-\beta\tfrac{1}{2}\boldsymbol{p}^T M^{-1}\boldsymbol{p} - \beta_c U\left(\boldsymbol{q}\right)\right]. \tag{4.2}$$

Observe that $Z(\beta)$ is used to denote a normalization constant parameterized by the reciprocal physical temperature β, i.e. for each $\beta \in \mathbb{R}_+$ there exists a corresponding $Z(\beta)$ defined as,

$$\begin{aligned} Z(\beta) &= \int_{\beta_{\min}}^{\beta_{\max}} \int_{\Omega\times\mathbb{R}^n} \rho_\beta\left(\boldsymbol{q},\boldsymbol{p},\beta_c\right) \mathrm{d}\boldsymbol{q}\,\mathrm{d}\boldsymbol{p}\,\mathrm{d}\beta_c, \\ &= (2\pi\beta)^{n/2}\left(\det M\right)^{1/2} \int_{\beta_{\min}}^{\beta_{\max}} \omega(\beta_c) Z_{\boldsymbol{q}}(\beta_c)\,\mathrm{d}\beta_c. \end{aligned} \tag{4.3}$$

Analogously $Z_{\boldsymbol{q}}(\beta_c)$ is used to make clear the parametric dependence on β_c in the configurational part of this partition function,

$$Z_{\boldsymbol{q}}(\beta) = \int_\Omega \exp[-\beta U(\boldsymbol{q})]\,\mathrm{d}\boldsymbol{q}\ . \tag{4.4}$$

There are a variety of different ergodic stochastic dynamics formulations that can be used to sample the density $\rho_\beta\left(\boldsymbol{q},\boldsymbol{p},\beta_c\right)$. Here we use the following extended Langevin dynamics,

$$\begin{aligned} \mathrm{d}\boldsymbol{q} &= M^{-1}\boldsymbol{p}\,\mathrm{d}t, \\ \mathrm{d}\boldsymbol{p} &= -\beta^{-1}\beta_c\nabla_{\boldsymbol{q}}U(\boldsymbol{q})\,\mathrm{d}t - \gamma M^{-1}\boldsymbol{p}\,\mathrm{d}t + \sigma_{\boldsymbol{p}}\,\mathrm{d}\boldsymbol{W}_{\boldsymbol{p}}, \\ \mathrm{d}\beta_c &= -\beta^{-1}U\left(\boldsymbol{q}\right)\varepsilon^{-1}\,\mathrm{d}t + \beta^{-1}\omega^{-1}(\beta_c)\omega'(\beta_c)\varepsilon^{-1}\,\mathrm{d}t + \sqrt{2\beta^{-1}}\sqrt{\varepsilon^{-1}}\,\mathrm{d}W_{\beta_c}. \end{aligned} \tag{4.5}$$

The noise satisfies $\sigma_{\boldsymbol{p}}^T\sigma_{\boldsymbol{p}} = 2\gamma\beta^{-1}$ in these equations and $\mathrm{d}\boldsymbol{W}_{\boldsymbol{p}}$, $\mathrm{d}W_{\beta_c}$ are a standard d-dimensional and scalar Wiener processes respectively. A time scaling parameter ε is introduced to control the time evolution of the temperature.

As $\varepsilon \to 0$ the temperature process is accelerated. In this limit it is identical to infinitely fast temperature switching since the $(\boldsymbol{q},\boldsymbol{p})$ dynamics evolves on a $\mathcal{O}\left(1\right)$ timescale whilst β_c on $\mathcal{O}\left(\varepsilon\right)$. Intuitively this implies that the extended dynamics $(\boldsymbol{q},\boldsymbol{p},\beta_c)$ is replaced by $(\boldsymbol{q},\boldsymbol{p},\mathbb{E}\left[\beta_c|\left(\boldsymbol{q},\boldsymbol{p}\right)\right])$, i.e. for any state $(\boldsymbol{q},\boldsymbol{p})$ the instant temperature in (4.5) is replaced by the average temperature observed for this state. Note that in this limit, an implementation of the β_c process is not necessary.

4.1.2 Averaged Equations of Motion

In this section we derive the averaged equations of motion, $\varepsilon \to 0$, using a stochastic homogenization argument closely following the treatment of Pavliotis and Stuart [83]. Secondly we justify the improved efficiency over (4.5) using a large deviation argument inspired by Dupuis et al in [80, 32].

The initial step in this process is to study a perturbation expansion in ε for the backward Kolmogorov equation. This results in a separation of timescales where the fast dynamics on β_c is studied for a fixed state $(\boldsymbol{q},\boldsymbol{p})$. Similarly, the slow dynamics on $\Omega\times\mathbb{R}^n$ can be studied for an equilibrium distribution of β_c resulting in reduced dynamics for $(\boldsymbol{q},\boldsymbol{p})$.

Define the generators of these two processes, i.e. from (4.5) we have,

$$\begin{aligned}\mathcal{L}_{\boldsymbol{q}} &= M^{-1}\boldsymbol{p}\cdot\nabla_{\boldsymbol{q}},\\ \mathcal{L}_{\boldsymbol{p}} &= -\beta^{-1}\beta_c\nabla_{\boldsymbol{q}}U(\boldsymbol{q})\cdot\nabla_{\boldsymbol{p}}+\gamma\left(-M^{-1}\boldsymbol{p}\cdot\nabla_{\boldsymbol{p}}+\beta^{-1}\Delta_{\boldsymbol{p}}\right),\\ \mathcal{L}_{\beta_c} &= \beta^{-1}\left(U(\boldsymbol{q})-\omega^{-1}(\beta_c)\omega'(\beta_c)\right)\partial_{\beta_c}+\beta^{-1}\partial_{\beta_c}^2.\end{aligned}\tag{4.6}$$

It is well established that (4.5) is ergodic with respect to the density (4.2). In turn, this implies that the reciprocal temperature dynamics, β_c, is ergodic for a fixed $(\boldsymbol{q},\boldsymbol{p})$ with respect to,

$$\rho(\beta_c\,;\boldsymbol{q},\boldsymbol{p}) = Z^{-1}(\boldsymbol{q},\boldsymbol{p})\,\omega(\beta_c)\exp\left[-\beta_c U(\boldsymbol{q})\right].\tag{4.7}$$

In this density, $(\boldsymbol{q},\boldsymbol{p})$ appear as parameters where $Z(\boldsymbol{q},\boldsymbol{p})$ is the correct normalization constant. The fast temperature dynamics, $\mathcal{L}_{\beta_c}$, for fixed $(\boldsymbol{q},\boldsymbol{p})$ satisfies,

$$\begin{aligned}\mathcal{L}_{\beta_c}1(\boldsymbol{q},\boldsymbol{p}) &= 0,\\ \mathcal{L}_{\beta_c}^{\dagger}\rho(\beta_c\,;\boldsymbol{q},\boldsymbol{p}) &= 0,\quad \text{for } (\boldsymbol{q},\boldsymbol{p}) \text{ fixed.}\end{aligned}\tag{4.8}$$

Here $1(\boldsymbol{q},\boldsymbol{p})$ denotes constants in $\Omega\times\mathbb{R}^n$.

To study the backward equation for (4.5) we introduce the irrelevant observable denoted as $\varphi(\boldsymbol{q},\boldsymbol{p},\beta_c,t)$ where,

$$\begin{aligned}&\varphi(\boldsymbol{q},\boldsymbol{p},\beta_c,t)\\ &\qquad=\mathbb{E}\left[\phi(\boldsymbol{q}(t),\boldsymbol{p}(t),\beta_c(t))\,|\boldsymbol{q}(0)=\boldsymbol{q},\boldsymbol{p}(0)=\boldsymbol{p},\beta_c(0)=\beta_c\right].\end{aligned}\tag{4.9}$$

The function, $\varphi\in\mathcal{D}(\mathcal{L})$ satisfies the backward equation whose generator we compose as the sum,

$$\partial_t\varphi=\mathcal{L}\varphi=\mathcal{L}_{\boldsymbol{q}}\varphi+\mathcal{L}_{\boldsymbol{p}}\varphi+\varepsilon^{-1}\mathcal{L}_{\beta_c}\varphi.\tag{4.10}$$

Here the time-scale separation between the temperature and physical coordinates is evident. Note that there is no dependence on $\boldsymbol{p}$ in $\mathcal{L}_{\beta_c}$ whereas $\boldsymbol{q}$ occurs as a parameter and is a result of restricting the tempering only in the configurational energy. Changing this perspective, to also include tempering of the kinetic energy would affect the coupling in the reciprocal temperature dynamics (4.5), with appropriate adjustments to the analysis.

It is straightforward to formally derive a multiscale expansion in $\varphi=\varphi_0+\varepsilon\varphi_1+\mathcal{O}\left(\varepsilon^2\right)$ to obtain,

$$\begin{aligned}\mathcal{O}\left(\varepsilon^{-1}\right)\quad &\mathcal{L}_{\beta_c}\varphi_0=0,\\ \mathcal{O}(1)\quad &\mathcal{L}_{\beta_c}\varphi_1=\partial_t\varphi_0-\left(\mathcal{L}_{\boldsymbol{q}}+\mathcal{L}_{\boldsymbol{p}}\right)\varphi_0\,.\end{aligned}\tag{4.11}$$

Using the ergodic properties in the first equation of (4.8) together with the $\mathcal{O}\left(\varepsilon^{-1}\right)$

equation in (4.11) and (4.8) we integrate the $\mathcal{O}(1)$ equation in (4.11),

$$\begin{aligned}\int_{\beta_{\min}}^{\beta_{\max}} \rho(\beta_c;\boldsymbol{q},\boldsymbol{p})\mathcal{L}_{\beta_c}\varphi_1 \,\mathrm{d}\beta_c &= \int_{\beta_{\min}}^{\beta_{\max}} \rho(\beta_c;\boldsymbol{q},\boldsymbol{p}) \left(\partial_t\varphi_0 - \left(\mathcal{L}_{\boldsymbol{q}} + \mathcal{L}_{\boldsymbol{p}}\right)\varphi_0\right) \mathrm{d}\beta_c \\ &= 0 \ .\end{aligned} \tag{4.12}$$

From this we obtain the new generator,

$$\partial_t\varphi_0 = \int_{\beta_{\min}}^{\beta_{\max}} \rho(\beta_c;\boldsymbol{q},\boldsymbol{p}) \left(\mathcal{L}_{\boldsymbol{q}} + \mathcal{L}_{\boldsymbol{p}}\right)\varphi_0 \,\mathrm{d}\beta_c = \bar{\mathcal{L}}\varphi_0, \tag{4.13}$$

and determines the averaged dynamics in $\Omega \times \mathbb{R}^n$. Invoking: $\int \rho(\beta_c;\boldsymbol{q},\boldsymbol{p})\,\mathrm{d}\beta_c = 1$, and exploiting the unique appearance of β_c in $\mathcal{L}_{\boldsymbol{p}}$, the generator can be simplified. To be more precise it is given as,

$$\begin{aligned}\bar{\mathcal{L}} = M^{-1}\boldsymbol{p}\cdot\nabla_{\boldsymbol{q}} - \beta^{-1}\int_{\beta_{\min}}^{\beta_{\max}} \beta_c\rho\left(\beta_c;\boldsymbol{q},\boldsymbol{p}\right)\mathrm{d}\beta_c\nabla_{\boldsymbol{q}}U\left(\boldsymbol{q}\right)\cdot\nabla_{\boldsymbol{p}} \\ + \gamma\left(-M^{-1}\boldsymbol{p}\cdot\nabla_{\boldsymbol{p}} + \beta^{-1}\Delta_{\boldsymbol{p}}\right),\end{aligned} \tag{4.14}$$

and the stochastic dynamics associated with the temperature averaged generator is,

$$\begin{aligned}\mathrm{d}\boldsymbol{q} &= M^{-1}\boldsymbol{p}\,\mathrm{d}t, \\ \mathrm{d}\boldsymbol{p} &= -\nabla_{\boldsymbol{q}}\left(-\beta^{-1}\log\int_{\beta_{\min}}^{\beta_{\max}} \omega(\beta_c)\exp\left[-\beta_c U\left(\boldsymbol{q}\right)\right]\mathrm{d}\beta_c\right)\mathrm{d}t \\ &\quad - \gamma M^{-1}\boldsymbol{p}\,\mathrm{d}t + \sigma_{\boldsymbol{p}}\,\mathrm{d}\boldsymbol{W}_{\boldsymbol{p}}.\end{aligned} \tag{4.15}$$

Here, $\sigma_{\boldsymbol{p}}^T\sigma_{\boldsymbol{p}} = 2\gamma\beta^{-1}$ and the invariant density that this dynamics is ergodic with respect to is given by,

$$\bar{\rho}_\beta\left(\boldsymbol{q},\boldsymbol{p}\right) = \bar{Z}^{-1}\exp\left[-\beta\left(\tfrac{1}{2}\boldsymbol{p}^T M^{-1}\boldsymbol{p} + \bar{U}(\boldsymbol{q})\right)\right] \ , \tag{4.16}$$

where $\bar{U}\left(\boldsymbol{q}\right)$ is given by equation (4.1) and the normalization constant is,

$$\bar{Z} = \int_{\beta_{\min}}^{\beta_{\max}}\int_{\mathbb{R}^n}\int_{\Omega} \omega(\beta_c)\exp\left[-\beta\tfrac{1}{2}\boldsymbol{p}^T M^{-1}\boldsymbol{p} - \beta_c U(\boldsymbol{q})\right]\mathrm{d}\boldsymbol{q}\,\mathrm{d}\boldsymbol{p}\,\mathrm{d}\beta_c. \tag{4.17}$$

It is clear from the above derivation that ST with infinitely fast temperature evolution, $\varepsilon \to 0$, is equivalent to the averaged dynamics (4.15). Similarly, this same limit is identical to sampling of an averaged potential given by $\bar{U}\left(\boldsymbol{q}\right)$ resulting in a rejection free algorithm that is simple to implement.

In conclusion, instead of sampling each temperature individually, as is done in standard implementations of ST, the equations in (4.15) sample an averaged potential $\bar{U}$, where the entire temperature range $[\beta_{\min}, \beta_{\min}]$ has been averaged over. Heuristically, this means that the samples drawn from this ensemble experience every temperature simultaneously.

4.1.3 Infinite Switch Limit

The above derivation shows that the dynamics in the limit $\varepsilon \to 0$ can be expressed in an averaged sense, where the temperature process is eliminated. In this section we use a large deviation argument similar to Dupuis et al. [32] that proves that (4.15) is more

efficient than (4.5). The process is straightforward and relies on the same additive form that we exploited in (4.10) that also generates a similar additive structure at the rate function level. This rate function measures the deviation of the empirical measure,

$$\lambda_{\mathcal{T}}\left(\boldsymbol{q},\boldsymbol{p},\beta_c\right)=\frac{1}{\mathcal{T}}\int_0^{\mathcal{T}}\delta(\boldsymbol{q}-\boldsymbol{q}(t))\delta(\boldsymbol{p}-\boldsymbol{p}(t))\delta(\beta_c-\beta_c(t))\,\mathrm{d}t, \tag{4.18}$$

in the asymptotic limit as $\mathcal{T}\to\infty$ to the measure,

$$\mu_\beta(\mathrm{d}\boldsymbol{q},\mathrm{d}\boldsymbol{p},\mathrm{d}\beta_c)=\rho_\beta(\boldsymbol{q},\boldsymbol{p},\beta_c)\,\mathrm{d}\boldsymbol{q}\,\mathrm{d}\boldsymbol{p}\,\mathrm{d}\beta_c, \tag{4.19}$$

where ρ_β is given by (4.2). In this asymptotic limit, a larger rate function indicates a smaller probability that the empirical measure $\lambda_{\mathcal{T}}$ deviates from the target measure μ_β. By showing that this rate function is monotonically increasing in ε^{-1} and we conclude that $\varepsilon\to 0$ is optimal for (4.5).

Theorem 2 (Optimality of the Infinite Switch Dynamics)**.** *Consider a process $(\boldsymbol{q},\boldsymbol{p},\beta_c)$ evolving according to the dynamics given by* (4.5)*. Let $\{\lambda_T\}_{T\geq 0}$ be a family probability measures in $\Gamma\subseteq P\left(\mathbb{R}^n\times\mathbb{R}^n\times(0,\infty)\right)$ with smooth density and absolutely continuous w.r.t. Lebesgue. The large deviation rate, indicating that the empirical measure λ_T is different from $\mu\in\Gamma$ is obtained from large deviation theory which ensures the existence of non negative rate functions satisfying,*

$$\lim_{\mathcal{T}\to\infty}\frac{1}{\mathcal{T}}\log\mathbb{P}\left(\lambda_T\approx\mu\right)=-I_\varepsilon\left(\mu\right)=-\left(J\left(\mu\right)+\varepsilon^{-1}\tilde{J}(\mu)\right). \tag{4.20}$$

Here $J(\mu)$ is the rate function in the slow $(\boldsymbol{q},\boldsymbol{p})$ dynamics and $\tilde{J}$ indicates the rate of decay in the fast β_c dynamics. The limit of this rate function satisfies,

$$I^\infty\left(\mu\right)=\lim_{\varepsilon\to 0}I_\varepsilon(\mu)=\begin{cases}J\left(\mu\right) & \textit{iff } \mu=\mu_\beta\\ \infty & \textit{otherwise}\end{cases}, \tag{4.21}$$

with μ_β the measure with density (4.2)*. The limit rate $I^\infty(\nu)$ is therefore optimal and the dynamics with this rate is given by* (4.15)*.*

Proof. According to Donsker-Varadhan theory [76], in the limit as $\mathcal{T}\to\infty$ the empirical measure (4.18) satisfies a large deviation principle,

$$\lim_{\mathcal{T}\to\infty}\frac{1}{\mathcal{T}}\log\mathbb{P}\left(\lambda_T\approx\mu\right)=-I_\varepsilon(\mu). \tag{4.22}$$

Define $f=\mathrm{d}\mu/\,\mathrm{d}\mu_{\beta_c}$ where we assume $\mu\ll\mu_{\beta_c}$. Then the rate in (4.22) is given by,

$$I_\varepsilon\left(\mu\right)=\sup_{f\in C^0_{f>0}(\mathbb{R}^n\times\mathbb{R}^n\times(0,\infty))}\int\frac{1}{f}\left(\mathcal{L}f\right)\mu(\mathrm{d}\boldsymbol{q},\mathrm{d}\boldsymbol{p},\mathrm{d}\beta_c). \tag{4.23}$$

Here, $\mathcal{L}$ is given by (4.10). Then by following Dupuis et. al. [32] and integration by

parts once, we find the rate functional as,

$$\begin{aligned}\tilde{J}(\mu) &= \int \frac{1}{f}\left(\mathcal{L}_{\beta_c} f\right) \mathrm{d}\mu(\mathrm{d}\boldsymbol{q}, \mathrm{d}\boldsymbol{p}, \mathrm{d}\beta_c), \\ &= \beta^{-1} \int \frac{1}{f}\left(U(q) - \partial_{\beta_c} \log \omega(\beta_c)\right) \partial_{\beta_c} f \, \mathrm{d}\mu(\mathrm{d}\boldsymbol{q}, \mathrm{d}\boldsymbol{p}, \mathrm{d}\beta_c) \\ &+ \beta^{-1} \int \frac{1}{f} \partial^2_{\beta_c} f \, \mathrm{d}\mu(\mathrm{d}\boldsymbol{q}, \mathrm{d}\boldsymbol{p}, \mathrm{d}\beta_c), \\ &= \frac{\beta^{-1}}{2} \int \frac{1}{f^2} |\partial_{\beta_c} f|^2 \, \mathrm{d}\mu(\mathrm{d}\boldsymbol{q}, \mathrm{d}\boldsymbol{p}, \mathrm{d}\beta_c).\end{aligned} \tag{4.24}$$

This implies that (4.22) is a monotone increasing function in ε^{-1} and this concludes the proof. □

4.1.4 Estimation of Canonical Expectations

The dynamics (4.15) is ergodic w.r.t. the measure with density (4.16). We are often interested in some observable $\varphi(\boldsymbol{q}, \boldsymbol{p})$ averaged w.r.t. the canonical measure at some fixed target temperature $\beta_i \in [\beta_{\min}, \beta_{\max}]$, with density given by:

$$\rho_{\beta_i}(\boldsymbol{q}, \boldsymbol{p}) = Z_{\beta_i}^{-1} \exp\left[-\beta_i H(\boldsymbol{q}, \boldsymbol{p})\right]. \tag{4.25}$$

Here, the normalisation constant is,

$$Z_{\beta_i} = \int \exp\left[-\beta_i H(\boldsymbol{q}, \boldsymbol{p})\right] \mathrm{d}\boldsymbol{q}\mathrm{d}\boldsymbol{p}. \tag{4.26}$$

The aim of this short section is to describe how the dynamics in (4.15), with invariant measure with density (4.16), can be used to determine averages of the form,

$$\mathbb{E}_{\beta_i}[\varphi] = \int \varphi(\boldsymbol{q}, \boldsymbol{p}) \rho_{\beta_i}(\boldsymbol{q}, \boldsymbol{p}) \, \mathrm{d}\boldsymbol{q}\mathrm{d}\boldsymbol{p}. \tag{4.27}$$

A straightforward calculation reveals that this can be achieved using importance weights of the form,

$$W_{\beta_i}(\boldsymbol{q}) = \frac{\bar{Z}}{Z_{\beta_i}} \left(\int \omega(\beta_c) e^{-(\beta_c - \beta_i) U(\boldsymbol{q})} \, \mathrm{d}\beta_c \right)^{-1}, \tag{4.28}$$

and (4.27) is given by,

$$\mathbb{E}_{\beta_i}[\varphi] = \int \varphi(\boldsymbol{q}, \boldsymbol{p}) \, W_{\beta_i}(\boldsymbol{q}) \, \bar{\rho}_{\beta}(\boldsymbol{q}, \boldsymbol{p}) \, \mathrm{d}\boldsymbol{q}\mathrm{d}\boldsymbol{p}. \tag{4.29}$$

Invoking ergodicity of (4.15) with respect to (4.16) and using (4.29), averages at any temperature $\beta_i \in [\beta_{\min}, \beta_{\max}]$ can be found using the weights defined as (4.28).

Note that (4.28) contains the in general unknown ratio $\bar{Z}/Z_{\beta_i}$, which can be found on-the-fly. This was first discussed by Carlson et. al [84]. In Section 4.1.6 we discuss how this ratio can be learned whilst simultaneously adjusting the weights. In essence, this boils down to utilising (4.27) with an identity observable and allows us to re-arrange (4.29) to find the ratio: $Z_{\beta_i}/\bar{Z}$.

4.1.5 Plausibility Argument for the Choice of Temperature Weights

In standard simulated tempering, the choice $\omega(\beta_c) = Z_{\boldsymbol{q}}^{-1}(\beta_c)$ – given in equation (4.4) – is justified as it leads to a uniform distribution in temperature. This is deemed advantageous as it means that all temperature levels are explored equally. A similar argument does not hold for the swapping free dynamics and it is necessary to provide a different argument. Below, we justify that this choice implies that the distribution of the dynamics explores energy levels uniformly.

Consider the distribution in energy of the measure with probability density given by (4.16), which we denote as,

$$\bar{\rho}(E) = \int \delta\left(E - U(\boldsymbol{q})\right) \bar{\rho}_\beta\left(\boldsymbol{q}, \boldsymbol{p}\right) \mathrm{d}\boldsymbol{q}\mathrm{d}\boldsymbol{p} = D(E) \frac{\int \omega(\beta_c) e^{-\beta_c E}}{\int Z_{\boldsymbol{q}}(\beta_c)\omega(\beta_c)\,\mathrm{d}\beta_c}. \tag{4.30}$$

Here the density of states is defined as,

$$D(E) = \int_\Omega \delta(U(\boldsymbol{q}) - E)\mathrm{d}\boldsymbol{q}. \tag{4.31}$$

Assume without loss of generality that $U(\boldsymbol{q}) \geq 0$ for $\boldsymbol{q} \in \Omega$. The standard expression for the configurational partition function $Z_{\boldsymbol{q}}(\beta_c)$ in terms of the density of states $D(E)$ is,

$$Z_{\boldsymbol{q}}(\beta_c) = \int_0^\infty e^{-\beta_c E} D(E)\,\mathrm{d}E. \tag{4.32}$$

We introduce two thermodynamic quantities: micro canonical entropy $S(E)$ and the canonical free energy $F(\beta_c)$, which are respectively defined as,

$$S(E) = \log D(E), \qquad F(\beta_c) = -\log Z_{\boldsymbol{q}}(\beta_c). \tag{4.33}$$

Using these definitions we rewrite (4.32) as,

$$F(\beta_c) = -\log \int_0^\infty e^{-\beta_c E + S(E)}\,\mathrm{d}E. \tag{4.34}$$

In the large system size limit we assume that the integral in (4.34) can be estimated asymptotically by the Laplace method, which implies that,

$$F(\beta_c) \sim \min_{E \geq 0} \left[\beta_c E - S(E)\right]. \tag{4.35}$$

Here we use $\sim$ to denote that the ratio of both sides tend to 1 in the thermodynamic limit i.e. the dimension goes to infinity. This implies that the free energy is the Legendre-Fenchel transform of the entropy and via the involution property of this transform we define the minimiser,

$$\min_{\beta_c \in [\beta_{\min}, \beta_{\max}]} \left[\beta_c E - F(\beta_c)\right] \sim S_\star(E). \tag{4.36}$$

The minimiser $S_\star(E)$ is the concave envelope of the entropy $S(E)$, and is uniquely defined if the derivative $\partial_{\beta_c} F(\beta_c)$ is continuous on $[\beta_{\min}, \beta_{\max}]$. The relation (4.36) can also be interpreted as,

$$e^{-S_\star(E)} \asymp \int_0^\infty e^{-\beta_c E + F(\beta_c)} \mathbb{1}_{[\beta_{\min}, \beta_{\max}]}(\beta_c)\,\mathrm{d}\beta_c = \int_{\beta_{\min}}^{\beta_{\max}} e^{-\beta_c E} Z_{\boldsymbol{q}}^{-1}(\beta_c)\,\mathrm{d}\beta_c, \tag{4.37}$$

where we use (4.33) in the final equality. Again, $\asymp$ is used to denote the thermodynamic limit where the logarithms of both sides tend to 1 as $n \to \infty$ i.e. large system limit. Combining (4.30) with (4.37) fixing $\omega(\beta_c) = Z_{\boldsymbol{q}}^{-1}(\beta_c)$, we find that,

$$\bar{\rho}(E) = D(E) \int_{\beta_{\min}}^{\beta_{\max}} Z_{\boldsymbol{q}}^{-1}(\beta_c) e^{-\beta_c E} \, \mathrm{d}\beta_c \asymp e^{S(E) - S_\star(E)}. \tag{4.38}$$

From this, we conclude that $\bar{\rho}(E) \asymp 1$ if $S(E) = S_\star(E)$, which is true if $S(E)$ is also convex and coincides with the minimiser $S_\star(E)$, and the energy distribution explored by the dynamics given by (4.15) is uniform.

The case when the minimiser $S_\star(E)$ is not unique arises when there is a value $\beta_{\text{critical}} \in [\beta_{\min}, \beta_{\max}]$, where the derivative of the free energy $\partial_{\beta_c} F(\beta_c)$ is not continuous. This indicates that the system is going through a first order phase transition at β_{critical} and we conclude that: $S(E) \neq S_\star(E)$. In this case β_{critical} is a bifurcation point of (4.36) and we have instead two branches,

$$\min_{\beta_c \in [\beta_{\min}, \beta_{\text{critical}})} [\beta_c E - F(\beta_c)] \sim S_\star^-(E), \quad \min_{\beta_c \in (\beta_{\text{critical}}, \beta_{\max}]} [\beta_c E - F(\beta_c)] \sim S_\star^+(E). \tag{4.39}$$

This means that the energy cannot be sampled uniformly and it is unclear whether the method provides an improvement in sampling the performance. We comment on these phenomena more closely in Chapter 6.

4.1.6 Adaptive Learning of Temperature Weights

In Section 4.1.5 above, we concluded that it is possible to sample uniformly in energy if $\omega(\beta_c) \propto Z^{-1}(\beta_c)$. In general, the partition function is unknown for all $\beta_i \in [\beta_{\min}, \beta_{\max}]$ and we must learn it using the estimator,

$$\frac{Z_{\beta_i}}{\bar{Z}} = \lim_{\mathcal{T} \to \infty} \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} \left(\int \omega(\beta_c) e^{-(\beta_c - \beta_i) U(\boldsymbol{q}(s))} \, \mathrm{d}\beta_c \right)^{-1} \mathrm{d}s. \tag{4.40}$$

This equation is derived from (4.29) with $\varphi = \mathbb{1}$. Let $\beta_i \in [\beta_{\min}, \beta_{\max}]$ be the target temperature for which we wish to determine the partition function. Introduce the two dynamical quantities $z(\beta_i, t)$ and $\omega(\beta_i, t)$. Assume that in the ergodic limit this dynamical quantity is proportional to the partition function i.e. $\lim_{t\to\infty} z(\beta_i, t) \propto Z_{\beta_i}$. Our goal is to define a recurrence relation such that also $\lim_{t\to\infty} \omega(\beta_i, t) \propto Z_{\beta_i}^{-1}$.

The algorithm is straightforward and consists of two parts,

- Define,
$$z(\beta_i, t) = \frac{1}{t} \int_0^t \left(\int \omega(\beta_c, s) e^{-(\beta_c - \beta_i) U(\boldsymbol{q}(s))} \, \mathrm{d}\beta_c \right)^{-1} \mathrm{d}s. \tag{4.41}$$

- Let,
$$\int_{\beta_{\min}}^{\beta_{\max}} \omega(\beta_c, t) \, \mathrm{d}\beta_c = 1, \quad \forall t \geq 0, \tag{4.42}$$
where for all $\beta_i \in [\beta_{\min}, \beta_{\max}]$,
$$\tau \partial_t \omega(\beta_i, t) = z^{-1}(\beta_i, t) - \lambda(t) \omega(\beta_i, t), \quad \text{with} \quad \lambda(t) = \int_{\beta_{\min}}^{\beta_{\max}} z^{-1}(\beta_c, t) \, \mathrm{d}\beta_c. \tag{4.43}$$

Here $\tau > 0$ is a timescale on which the weights are updated and the constraint (4.42) is enforced by the Lagrange multiplier $\lambda(t)$ in (4.43).

The dynamics given in (4.15) is updated to include the dynamical estimate $\omega(\beta_c, t)$ such that the new dynamics that we consider is,

$$\begin{aligned} \mathrm{d}\boldsymbol{q} &= M^{-1}\boldsymbol{p}\,\mathrm{d}t, \\ \mathrm{d}\boldsymbol{p} &= -\nabla_{\boldsymbol{q}}\left(-\beta^{-1}\log\int_{\beta_{\min}}^{\beta_{\max}}\omega(\beta_c,t)\exp\left[-\beta_c U\left(\boldsymbol{q}\right)\right]\mathrm{d}\beta_c\right)\mathrm{d}t \\ &\quad -\gamma M^{-1}\boldsymbol{p}\,\mathrm{d}t + \sigma_{\boldsymbol{p}}\,\mathrm{d}\boldsymbol{W}_{\boldsymbol{p}}. \end{aligned} \tag{4.44}$$

Note that all the parameters are identical to before, but the weight is now dynamical as indicated by the $\omega(\beta_c, t)$ term. To show that these equations are justified, we need to show that the dynamics (4.43) has a fixed point and that $\omega(\beta_c, t)$ converges to it.

Setting $\tau = \infty$ in (4.43) fixes $\omega(\beta_c, t) = \omega(\beta_c, 0)$ for all t and implies that,

$$\begin{aligned} \lim_{t\to\infty} z(\beta_i, t) &= \lim_{t\to\infty}\frac{1}{t}\int_0^t\left(\int\omega(\beta_c,s)e^{-(\beta_c-\beta_i)U(\boldsymbol{q}(s))}\,\mathrm{d}\beta_c\right)^{-1}\mathrm{d}s \\ &= \frac{Z_{\boldsymbol{q}}(\beta_i)}{\int_{\beta_{\min}}^{\beta_{\max}} Z_{\boldsymbol{q}}(\beta_c)\omega(\beta_c,0)\,\mathrm{d}\beta_c}. \end{aligned} \tag{4.45}$$

Here we combined (4.41) with (4.4). Next, when $\tau < \infty$ assume that there exists a fixed point for (4.43) which we label $\omega_\infty(\beta_c)$. Using (4.43) we conclude that, together with the constraint (4.42), this fixed point satisfies,

$$\omega_\infty(\beta_i) = \frac{z_\infty^{-1}(\beta_i)}{\int_{\beta_{\min}}^{\beta_{\max}} z_\infty^{-1}(\beta_c)\,\mathrm{d}\beta_c}, \tag{4.46}$$

with,

$$z_\infty(\beta_i) = \frac{Z_{\boldsymbol{q}}(\beta_i)}{\int_{\beta_{\min}}^{\beta_{\max}} Z_{\boldsymbol{q}}(\beta_c)\omega_\infty(\beta_c)\,\mathrm{d}\beta_c}, \tag{4.47}$$

again using (4.41). We conclude that $\omega_\infty(\beta_i) \propto Z_{\boldsymbol{q}}^{-1}(\beta_i)$. In a final remark we highlight that for both cases: $\tau = \infty$ and $\tau < \infty$, the relative size of the partition functions are conserved in the ergodic limit since,

$$\frac{\lim_{t\to\infty} z(\beta_c, t)}{\lim_{t\to\infty} z(\beta_c', t)} = \frac{Z_{\boldsymbol{q}}(\beta_c)}{Z_{\boldsymbol{q}}(\beta_c')}. \tag{4.48}$$

This argument implies the existence of a fixed point, but not whether it is stable or not. In the following sections we demonstrate numerically that this scheme works.

4.2 Implementation details of the ISST algorithm

Let us now discuss the practical aspects of the ISST algorithm. For the purpose of discretizing the limiting equation (4.15), we suggest to use the second order "BAOAB"

Langevin scheme [16], of the form

$$\begin{aligned}
\boldsymbol{p}_{n+1/2} &= \boldsymbol{p}_n - \tfrac{1}{2}\Delta t\beta^{-1}\nabla\bar{U}(\boldsymbol{q}_n),\\
\boldsymbol{q}_{n+1/2} &= \boldsymbol{q}_n + \tfrac{1}{2}\Delta t M^{-1}\boldsymbol{p}_{n+1/2},\\
\hat{\boldsymbol{p}}_{n+1/2} &= e^{-\Delta t\gamma}\boldsymbol{p}_{n+1/2} + [\beta^{-1}(1-e^{-2\gamma\Delta t})m]^{1/2}\boldsymbol{\eta}_n,\\
\boldsymbol{q}_{n+1} &= \boldsymbol{q}_{n+1/2} + \tfrac{1}{2}\Delta t M^{-1}\hat{p}_{n+1/2},\\
\boldsymbol{p}_{n+1} &= \hat{p}_{n+1/2} - \tfrac{1}{2}\Delta t\beta^{-1}\nabla\bar{U}(\boldsymbol{q}_{n+1}),
\end{aligned} \tag{4.49}$$

where $(\boldsymbol{q}_n, \boldsymbol{p}_n)$ are the time-discretized approximations of $(\boldsymbol{q}(n\Delta t), \boldsymbol{p}(n\Delta t))$, Δt is the timestep, and $\eta_n \sim \mathcal{N}(0,1)$. This method is known to have low configurational sampling bias in comparison with other Langevin MD schemes [85].

In order to make the scheme above explicit, one needs to estimate the force in (4.15), i.e.. provide a scheme to evaluate $\nabla\bar{U}(\boldsymbol{q})$ where we let $\nabla\bar{U}(\boldsymbol{q}) = \bar{\beta}(U(\boldsymbol{q}))U(\boldsymbol{q})$. This involves addressing two issues: the first is how to estimate the 1–dimensional integral in (4.15) given the weights $\omega(t,\beta_c)$; the second is how to update the weights by discretizing the equations given in Sec. 4.1.6.

Regarding the first issue, any quadrature (numerical integration) method can in principle be used. However, since this quadrature rule is part of an iterative 'learning' strategy in which statistics are accumulated on-the-fly to update the weights, it is desirable to use a fixed set of nodes or grid points $\{\beta_i\}_{1\le i\le M}$, so that the corresponding samples collected at earlier stages remain relevant as the system is updated.

For a fixed number of nodes, the optimal choice of quadrature rule on a given interval $[\beta_{\min}, \beta_{\max}]$ in terms of accuracy is derived by placing the nodes at the roots of a suitably adjusted Legendre polynomial (Gauss-Legendre quadrature). Let the quadrature weight for node i be B_i and replace $\bar{\beta}(U(\boldsymbol{q}_n))$ in (4.15) by,

$$\hat{\beta}(U(\boldsymbol{q}_n)) = \frac{\sum_{i=1}^{M} B_i\beta_i\omega_{i,n}e^{-\beta_i U(\boldsymbol{q}_n)}}{\sum_{i=1}^{M} B_i\omega_{i,n}e^{-\beta_i U(\boldsymbol{q}_n)}}. \tag{4.50}$$

where $\omega_{i,n}$ is the current estimate of the weight at node i.

To obtain $\omega_{i,n}$ we use the following discrete recurrence relation consistent with (4.43). Given $\omega_{i,n}$, we find $\omega_{i,n+1}$ via,

$$\omega_{i,n+1} = \frac{\omega_{i,n+1}}{\sum_{j=1}^{M} B_j\omega_{j,n+1}}\ , \tag{4.51}$$

in which

$$\omega_{i,n+1} = \left(1-\tau^{-1}\Delta t\right)\omega_{i,n} + \tau^{-1}\Delta t\, z_{i,n}^{-1}, \tag{4.52}$$

with

$$z_{i,n} = \frac{1}{n}\sum_{m=1}^{n}\frac{e^{-\beta_i U(\boldsymbol{q}_m)}}{\sum_{j=1}^{M} B_j\omega_{j,m}e^{-\beta_j U(\boldsymbol{q}_m)}}. \tag{4.53}$$

Here, $\tau > 0$ is the time-scaling parameter introduced in Sec. 4.1.6. It can be checked that this recurrence relation preserves the constraint that for all $n \ge 0$,

$$\sum_{j=1}^{M} B_j\omega_{j,n} = 1\ , \tag{4.54}$$

and that (4.51) and (4.52) are consistent with (4.42). Note also that (4.53) can be

written in terms of the following iteration rule

$$z_{i,n} = \frac{1}{n} \frac{e^{-\beta_i U(\boldsymbol{q}_n)}}{\sum_{j=1}^{M} B_j \omega_{j,n} e^{-\beta_j U(\boldsymbol{q}_n)}} + \frac{n-1}{n} z_{i,n-1}. \tag{4.55}$$

This quantity gives a running estimate of the ratio in (4.45).

The discussion above makes apparent that ISST requires very few adjusting parameters besides the one already used in a vanilla MD code (i.e.. the parameters like Δt in (4.49)): The user is required to choose the temperature range $[\beta_{\min}, \beta_{\max}]$ and the scaling parameter τ. Other parameters, like the number and positions of the nodes for the temperature or the quadrature rule to be used, can be adjusted using standard practices in numerical quadrature, and they do not affect the overall efficiency of the method.

4.3 Numerical Experiments

In order to evaluate the performance of the discretization method described in the previous section we now present results from several numerical experiments on three test systems: the d–dimensional harmonic oscillator; the continuous Curie-Weiss model, a mean field version of the Ising model which displays a second-order phase transition in temperature; and the Alanine-12 molecule in vacuum, which displays conformational changes. On these examples we investigate (i) the influence of the number of quadrature points, M, used in the ISST algorithm when the weights ω_i are known; (ii) the convergence of (4.53), both when the weights are fixed to their initial (and non-optimal) value ($\tau \to \infty$ limit) and when these weights are adjusted towards their optimal values; and (iii) the effect of the choice of τ on the convergence of the weights ω_i, estimated using (4.51)-(4.53). We also compare the efficiency of ISST and standard ST on the last two examples.

4.3.1 Harmonic Oscillator

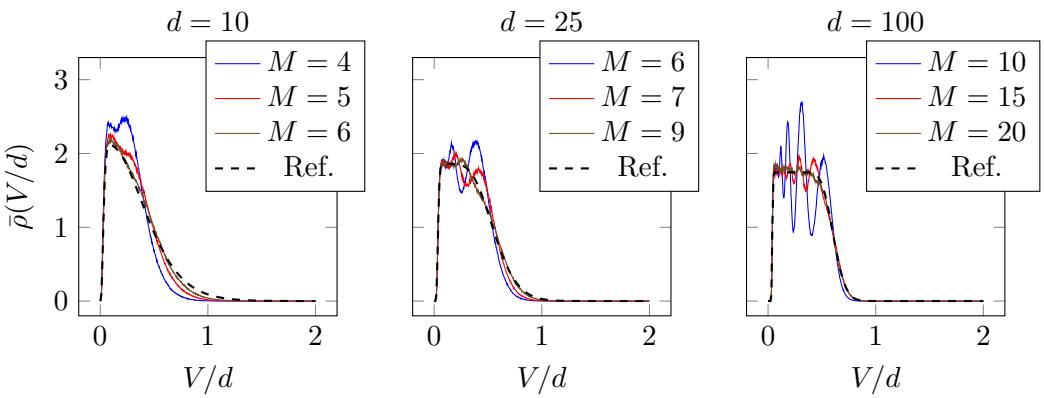

Figure 4.1: Behavior of $\bar{\rho}(E)$, (4.38) with $E = V/d$, for the harmonic oscillator. The results for different numbers of temperatures M were obtained by recording the energy of an ISST trajectory of length $N = 10^7$ steps using a histogram. The ISST weights are given by (4.62).

Consider a d-dimensional harmonic oscillator given by the quadratic potential in

$\mathbb{R}^d$,

$$V(\boldsymbol{q}) = \frac{1}{2}\sum_{j=1}^{d} \lambda_j q_j^2, \tag{4.56}$$

where $\{\lambda_j\}_{j=1}^d$ is a set of positive constants. The partition function $Z_{\boldsymbol{q}}(\beta)$ can be written explicitly as,

$$Z_{\boldsymbol{q}}(\beta) = A\beta^{-d/2} \qquad \text{with} \quad A = (2\pi)^{d/2}\prod_{j=1}^{d}\lambda_j^{-1/2}. \tag{4.57}$$

The goal is to perform simulations using (4.49) and (4.50) with some yet to be determined weights ω_i. As we showed in Section 4.1.5 the asymptotic optimal weight is $\omega(\beta) \propto Z_q^{-1}(\beta)$, which implies that $\omega_i \propto Z_q^{-1}(\beta_i)$. This leads to a log-asymptotically uniform energy in (4.38) i.e. $\bar{\rho}(E) \asymp 1$.

Using (4.57) we can write the density of states, for the potential given by (4.56) as,

$$D(E) = \frac{AE^{1-d/2}}{\Gamma(\frac{d}{2})}. \tag{4.58}$$

Additionally, since $D^{-1}(E)$ is completely monotonic for $d > 2$, the Hausdorff-Bernstein-Widder-theorem guarantees the existence of a measure $\mu(\beta_c)$ [31] such that,

$$D^{-1}(E) = \int_0^\infty e^{-\beta_c E}\,\mathrm{d}\mu(\beta_c). \tag{4.59}$$

It is straightforward to verify that this measure is

$$d\mu(\beta_c) = C^{-1}\beta_c^{d/2-2}d\beta_c, \qquad \text{with} \quad C = \frac{\Gamma(\frac{d}{2}-1)\,A}{\Gamma(\frac{d}{2})}. \tag{4.60}$$

With the knowledge of (4.59), it is easy to see that (4.38) requires that $\bar{\rho}(E) = 1$ as $d \to \infty$ if $\omega(\beta_c) \propto \beta_c^{d/2-2}$. This suggests that one could use,

$$\omega(\beta_c) \propto \begin{cases} \beta_c^{d/2-2}, & \beta_c \in [\beta_{\min}, \beta_{\max}] \\ 0, & \text{else} \end{cases} \tag{4.61}$$

The constant of proportionality is determined so as to satisfy (4.54), such that the explicit form of the M weights in (4.50) is defined by,

$$\omega_i = \beta_i^{d/2-2}\left(\sum_{j=1}^{M} B_j\beta_j^{d/2-2}\right)^{-1}. \tag{4.62}$$

In Figure 4.1 we show results for $d = 10,\ 25$ and 100, sampled using (4.49), (4.50) and (4.62) with a total trajectory length of $N = 10^7$ with $\Delta t = 0.1$ and $\beta_{\min} = 0.8$ and $\beta_{\max} = 12.5$. Each panel shows the convergence to the dashed reference (4.53) found using quadrature. We conduct experiments for three values of the dimension d, in which we vary the number of quadrature points (or reciprocal temperatures) M. The figure clearly illustrates the importance of choosing an appropriate number of points M, such that the observable of interest has a satisfactory support. Also note the dependence of

M on the dimension d, which is an entropic effect resulting from the dependence of the potential (4.56) on d.

Adaptive Weight Learning for the Harmonic Oscillator.

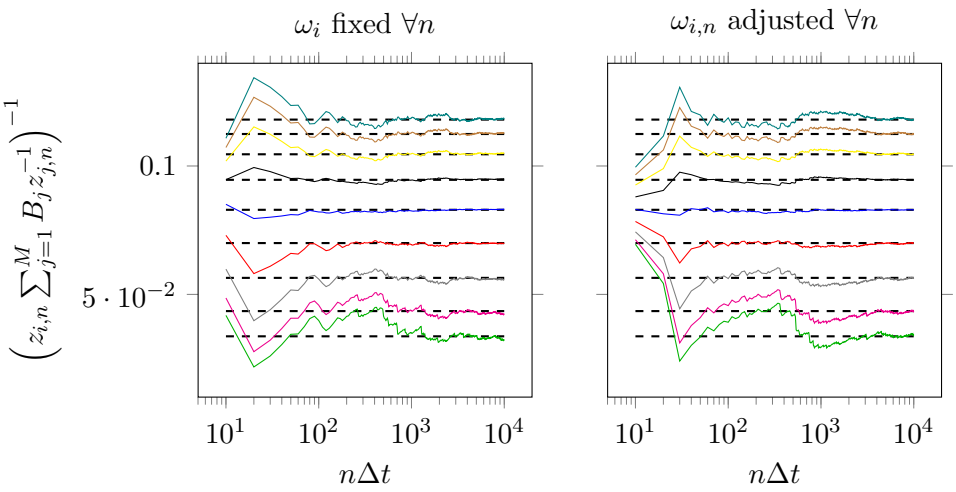

Figure 4.2: Reciprocal of (4.63) learned using $M = 10$ temperatures between $\beta_{\max} = 12.5$ and $\beta_{\min} = 0.8$ for (4.56) with $d = 1$ and $\Delta t = 0.01$ with $\tau = 1$ in (4.52). The black dashed lines show the asymptotic long-term average (4.64). In the left panel we keep the weight fixed for all simulation time and in the right panel we update the weights at every timestep.

In this section we check the convergence in time of quantity,

$$z_{i,n}\sum_{j=1}^{M} B_j z_{j,n}^{-1}, \tag{4.63}$$

to (4.47) with $z_{i,n}$ given by (4.53). This is a normalized version of the partition function whose inverse gives the optimal weight (2.1).

We perform a comparative experiment between two variants of the estimate (4.63). First we initialize the weights at $\omega_i \propto 1$, normalize according to (4.54) and fix these weights for a complete ISST simulation. Secondly, we instead initialize $\omega_{i,0} \propto 1$ and normalize according to (4.54), and adjust the weights at every timestep as described in Sec. 4.2.

In Figure 4.2 we show the results of these experiments using (4.56) with $d = 1$ and $M = 10$ reciprocal temperatures between $\beta_{\min} = 0.8$ and $\beta_{\max} = 12.5$. In the left panel we present the results of the first experiment described above, in which we fix the weights ω_i for all simulation time–as indicated by the title. To the right, we show the second experiment in which we adjust the weights at every timestep via (4.51), (4.52) and (4.54).

Both panels in Figure 4.2 show the reciprocal of (4.63) for all the M temperatures in color, whereas in dashed black we show the time asymptotic behaviour. The time-

asymptotic limit as $n \to \infty$ in (4.63) is:

$$\beta_i^{1/2} \left(\sum_{j=1}^{10} B_j \beta_j^{1/2} \right)^{-1} . \tag{4.64}$$

It is clear from Figure 4.2 that it is possible to learn ratios of the partition functions for a modest number of timesteps n, regardless of the value of ω_i. In practice one does not wish to fix the weights at some non-optimal value, as was done initially in this section, as this will most likely impede the sampling efficiency of the algorithm. Instead it is preferable to make use of the second approach, where one adjusts the weights continuously towards some optimum, as the simulation progresses.

Convergence of the Temperature Weights

The combination of the ISST Langevin scheme (4.49) and the adaptive weight-learning (4.51) results in a powerful, simple-to-implement sampling algorithm. In Sec. 4.1.6 we introduced a timescale parameter τ which adjusts the rate of weight learning in relation to the timestep in the ISST scheme. This section aims at exploring the choice of this parameter and its effect on the convergence of the weights.

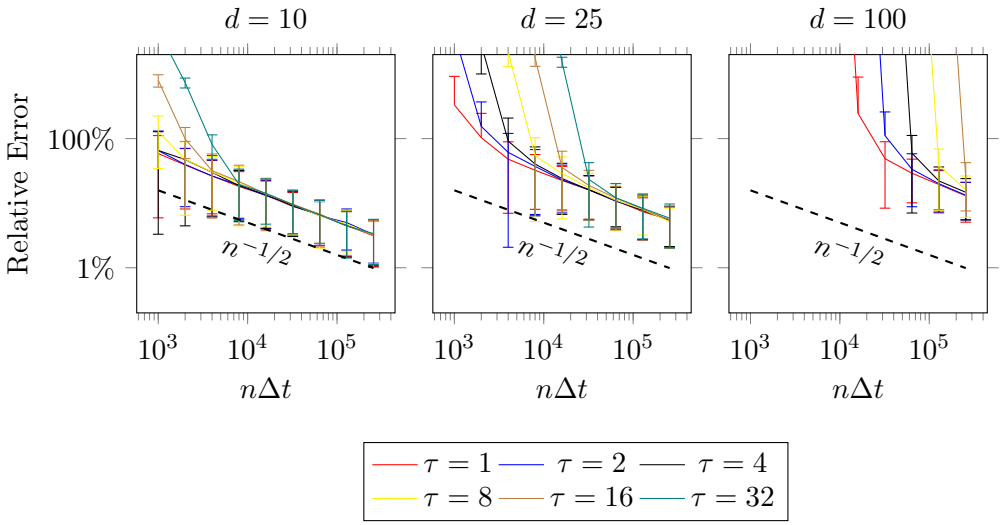

Figure 4.3: Convergence rate for a d-dimensional harmonic oscillator using $M = 10$ temperatures between $\beta_{\min} = 0.08$ and $\beta_{\max} = 12.5$ with $\Delta t = 0.1$ and a wide range of τ. As τ is increased the dynamics of the weight-learning algorithm slows down, which consequently slows down the convergence of the weights. The convergence rate $n^{-1/2}$ is the standard error decay of Monte Carlo averages. Each data point was calculated by averaging over 200 independent ISST trajectories and the error bars are the standard deviations associated with these averages.

We define the relative error as

$$\text{Rel.Error} = \sum_{M \geq i > 0} \frac{|\omega_{i,N} - \omega_{i,\infty}|}{|\omega_{i,\infty}|} \qquad \text{with} \qquad \omega_{i,\infty} = \beta_i^{d/2} \left(\sum_{j=1}^{M} B_j \beta_j^{d/2} \right)^{-1} , \tag{4.65}$$

where N refers to the last timestep of the simulation. We use (4.65) as a metric for

the accuracy of the approximations from (4.51), i.e. $\omega_{i,n}$ for $0 \leq n \leq N$. Working with $M = 10$ temperatures between $\beta_{\min} = 0.08$ and $\beta_{\max} = 12.5$ we perform experiments using (4.49) with $\Delta t = 0.1$ varying τ in (4.52). The results of this experiment with initial condition $\omega_{i,0} \propto 1$ for all i, are shown in Figure 4.3. In Sec. 4.1.6 we indicated that the fixed point of the learning scheme should be stable as $\tau \to \infty$ and we now observe that, at least in this example, it is stable even for moderate values of τ. In fact it appears that there is no advantage of using τ large and we see that its only effect, in the toy model, is to slow the convergence to the fixed point. In more complicated systems the choice of τ will be more critical.

The previous section implied that the adjustment scheme (4.51) for the weights $\omega_{i,n}$, should be dependent on the approximation of the ratio of partition functions $z_{i,n}$. Figure 4.3 makes it clear that the estimation of $z_{i,n}$ dominates the error when estimating the weights $\omega_{i,n}$. Consequently the observed $n^{-1/2}$ convergence with timestep is a result of this Monte Carlo averaging. Accuracy in $\omega_{i,n}$ can therefore only be gained by extending simulation time or alternatively by using accelerated stochastic approximation techniques.

4.3.2 Curie-Weiss Magnet

We next consider a continuous version of the Curie-Weiss magnet, i.e.. the mean field Ising model with K spins and potential

$$V_K(\theta_1, \ldots, \theta_K; b) = -\frac{1}{2K}\left(\sum_{i=1}^{K} \cos\theta_i\right)^2 - b\sum_{i=1}^{K}\cos\theta_i, \tag{4.66}$$

where $b \in \mathbb{R}$ is the intensity of the applied field. The Gibbs (canonical) density for this model is,

$$\varrho_K(\theta_1, \ldots, \theta_K, \beta, b) = \mathcal{Z}_K^{-1}(\beta, b)\exp\left[-\beta V_K(\theta_1, \ldots, \theta_K; b)\right], \tag{4.67}$$

where,

$$\mathcal{Z}_K(\beta, b) = \int_{[-\pi,\pi]^K} \exp\left[-\beta V_K(\theta_1, \ldots, \theta_K; b)\right] d\theta_1 \ldots \theta_K. \tag{4.68}$$

This system has similar thermodynamic properties to the standard Curie-Weiss magnet with discrete spins, but it is amenable to simulation by Langevin dynamics since the angles θ_i vary continuously. That is, we can simulate it in the context of ST in the infinite switch limit using (4.15) with $(\theta_1, \ldots, \theta_K)$ playing the role of q.

Thermodynamic properties and phase transition diagram

As in the standard Curie-Weiss magnet, the system with potential (4.66) displays phase transitions when β is varied with $b = 0$ fixed and when b is varied with β fixed above a critical value. To see why, and also to introduce a quantity that we will monitor in our numerical experiments, let us marginalize the Gibbs density (4.67) in the average magnetization m defined as

$$m = \frac{1}{K}\sum_{i=1}^{K}\cos\theta_i. \tag{4.69}$$

This marginalized density is given by

$$\rho_K(m,\beta,b) = \int_{[-\pi,\pi]^K} \varrho_K(\theta,\beta,b)\delta\left(m - \sum_{i=1}^{K}\cos\theta_i\right) d\theta_1 \ldots \theta_K. \tag{4.70}$$

A simple calculation shows that

$$\rho_K(m,\beta,b) = Z_K^{-1}(\beta,b)\exp\left[-\beta K F_K(m;\beta,b)\right], \tag{4.71}$$

where $Z_K(\beta,b) = \int_{-1}^{1} e^{-\beta K F_K(m;\beta,b)} dm$ and we introduced the (scaled) free energy $F_K(m;\beta,b)$ defined as

$$F_K(m;\beta,b) = V(m;b) - \beta^{-1} S_K(m) \tag{4.72}$$

with potential term

$$V(m;b) = -\tfrac{1}{2}m^2 - bm, \tag{4.73}$$

and entropic term

$$S_K(m) = K^{-1}\log \int_{[-\pi,\pi]^K} \delta\left(m - \sum_{i=1}^{K}\cos\theta_i\right) d\theta_1 \ldots \theta_K. \tag{4.74}$$

The marginalized density (4.71) and the free energy (4.72) can be used to analyze the properties of the system in thermodynamic limit when $K \to \infty$ and map out its phase transition diagram in this limit. In particular, we show next that $F_K(m;\beta,m)$ has a limit as $K \to \infty$ that has a single minimum at high temperature, but two minima at low temperature. Since $F_K(m;\beta,m)$ is scaled by K in (4.71), this implies that the density can become bimodal at low temperature, indicative of the presence of two strongly metastable states separated by a free energy barrier whose height is proportional to K.

The limiting free energy $F(m;\beta,b)$ is defined as

$$\begin{aligned} F(m;\beta,b) &= \lim_{K\to\infty} F_K(m;\beta,b) \\ &= -\tfrac{1}{2}m^2 - bm - \beta^{-1}\lim_{K\to\infty} S_K(m). \end{aligned} \tag{4.75}$$

To calculate the limit of the third (entropic) term, let us define $H(\lambda)$ via the Laplace transform of (4.74) through

$$\begin{aligned} e^{-KH(\lambda)} &= \int_{-1}^{1} e^{-K\lambda m + K S_K(m)} dm \\ &= \int_{[-\pi,\pi]^K} e^{-\lambda\sum_{i=1}^{K}\cos(\theta_i)} d\theta_1 \ldots \theta_K \\ &= \prod_{i=1}^{K}\int_{-\pi}^{\pi} e^{-\lambda\cos\theta_i} d\theta_i \\ &= \left(2\pi I_0(\lambda)\right)^K, \end{aligned} \tag{4.76}$$

where $I_0(\lambda)$ is a modified Bessel function. In the large K limit, $S(m)$ can be calculated

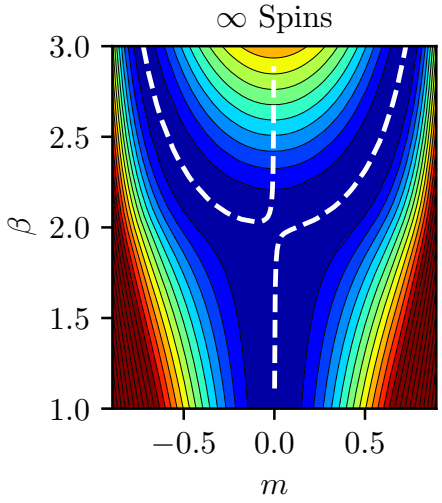

Figure 4.4: Limiting free energy $F(m;\beta,b)$ as $K\to\infty$ for (m,β) using a modest external force to bias the system, $b=0.001$. The averaged magnetization where $F'(m;\beta,b)=0$ is shown in white and the contour is the free energy surface (4.75).

from $H(\lambda)$ by Legendre transform

$$\begin{aligned} S(m) = \lim_{K\to\infty} S_K(m) &= \min_{\lambda}\{\lambda m - H(\lambda)\} \\ &= \min_{\lambda}\{\lambda m + \log I_0(\lambda)\} + \log(2\pi). \end{aligned} \tag{4.77}$$

The minimizer $\lambda(m)$ of (4.77) satisfies

$$m = -\frac{I_1(\lambda(m))}{I_0(\lambda(m))}, \tag{4.78}$$

which, upon inversion, offers a way to parametrically represent $S(m)$ using

$$S(m(\lambda)) = \lambda m(\lambda) + \log I_0(\lambda) + \log(2\pi), \qquad m(\lambda) = -\frac{I_1(\lambda)}{I_0(\lambda)}, \qquad \lambda\in\mathbb{R}. \tag{4.79}$$

Similarly we can represent $F(m;\beta,b)$ as

$$F(m(\lambda);\beta,b) = -\tfrac{1}{2}m^2(\lambda) - bm(\lambda) - \beta^{-1}S(m(\lambda)), \qquad m(\lambda) = -\frac{I_1(\lambda)}{I_0(\lambda)} \qquad \lambda\in\mathbb{R}. \tag{4.80}$$

In Figure 4.4 we show a contour plot of (4.75) as a function of m and β for fixed b obtained using this representation. Also shown is the location of the minima of $F(m;\beta,b)$ in the (m,β) plane at b fixed. These minima can also be expressed parametrically. Indeed, (4.79) implies that

$$S'(m(\lambda)) = \lambda, \tag{4.81}$$

which if we use it in $F'(m;\beta,b)=0$ to locate the minima of the free energy in the

(m, β) plane, indicates that they can be expressed parametrically as

$$\beta(\lambda) = \frac{\lambda}{m(\lambda) + b}, \qquad m(\lambda) = -\frac{I_1(\lambda)}{I_0(\lambda)}, \qquad \lambda \in \mathbb{R}. \tag{4.82}$$

The corresponding path gives the averaged magnetization as a function of β and is shown as a dashed line in Fig. 4.4 and was plotted using these formulae with $b = 0.001$. For values of β less than 2, the free energy is a single-well, and the averaged magnetization is approximately zero. For values of β above 2, the free energy becomes a double-well, and two metastable states with nonzero magnetization emerge.

If we consider the case $b = 0$, then by symmetry, $m = 0$ is a critical point of $F(m, \beta, b = 0)$ for all values of β, i.e.. $F'(0, \beta, b = 0) = 0$. By differentiating (4.81) in λ using the chain rule, we deduce that

$$S''(m(\lambda)) = 1/m'(\lambda) \tag{4.83}$$

which, if we evaluate it at $\lambda = 0$ using $m(\lambda = 0) = 0$ as well as $m'(\lambda = 0) = -\frac{1}{2}$ which follows from $m(\lambda) = -I_1(\lambda)/I_0(\lambda)$, indicates that

$$S''(0) = -2. \tag{4.84}$$

As a result

$$F''(0; \beta, b = 0) = -1 + 2\beta^{-1} \tag{4.85}$$

which means that $m = 0$ is a stable critical point of $F(m; \beta, b = 0)$ for $\beta < \beta_c = 2$, and an unstable critical point for $\beta > \beta_c = 2$, with a phase transition occurring at $\beta_c = 2$. A similar calculation can be performed when $b \neq 0$, but it is more involved since $m = 0$ is not a critical point in this case (hence we need to solve (4.82) numerically in λ to express the critical m as a function of β): in this case, the location of the global minimum of $F(m; \beta, b)$ varies continuously, so strictly speaking there is no phase transition.

It should be stressed that the phase transition observed when β is varied at $b = 0$ fixed is second order, i.e.. $G(\beta) = -K^{-1} \log Z_K(\beta, b = 0)$ is continuous with a continuous first order derivative in β at $\beta = \beta_c = 2$, but discontinuous in its second order derivative at that point. As a result the phase transition observed in the model above does not lead to difficulties of the kind discussed in Section 4.1.5: in particular it can be checked by direct calculation that $S_*(E) = S(E)$ (i.e.. the entropy $S(E)$ is concave down).

Sampling near or at the Phase Transition

In this section we use (4.49) with weights adjusted as in (4.51) to sample (4.71) over a range of temperatures, from *high*, for which (4.71) is unimodal, to *low*, for which it is bimodal. This is challenging for standard sampling methods because, as indicated by the results in Sec. 4.3.2, at low temperature the system has two metastable states separated by an energy barrier whose height scales linearly with K, as K increases.

The results of our experiments using varying numbers of spins are presented in the four panels of Figure 4.5. The minimum of the sampled free energy in the lower half of the magnetization range is shown in red, and the minimum of the upper half in blue. In dashed black we show the averaged magnetization (minimum of of the free energy (4.75)) in the $K \to \infty$ limit, as a reference guide. Each point in the collection of sampled minima was calculated by recording the average magnetization in a histogram. This was repeated for 20 independent ISST simulations, each of length $N = 10^5$ with

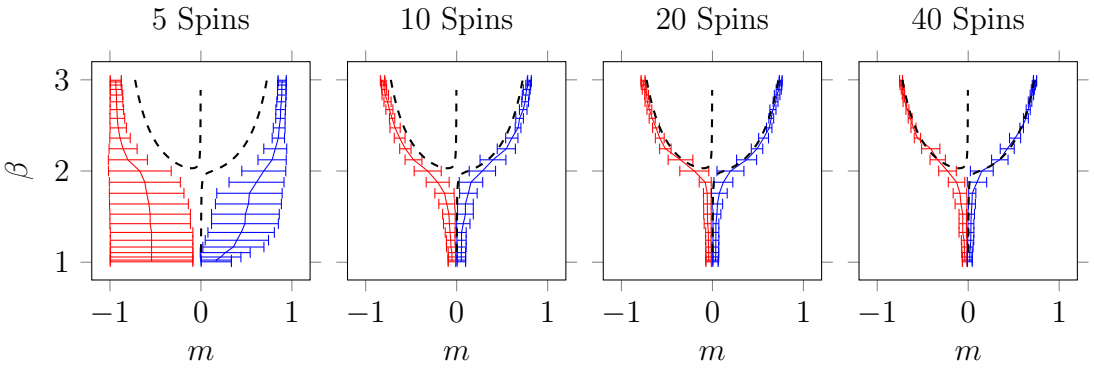

Figure 4.5: Minimum of the free energy vs the number of spins, in comparison to the theoretical minimum for $K \to \infty$ shown as dashed black. The different colour lines show the minimum in the upper and lower half, respectively. Each data point was calculated by averaging over 20 ISST simulations of length $N = 10^5$ with timestep $\Delta t = 0.1$ and $M = 25$. The minimum of the magnetization m was found by collecting points from the trajectories in a histogram with 200 bins, the minimum was then found in the upper (respectively lower)100 bins and are shown in red (and blue).

$\Delta t = 0.1$, whose average was used to find the minima. The error bars show the standard deviation of these 20 experiments.

We clearly observe in Figure 4.5 that the ISST algorithm encounters no difficulties in sampling the free energy surface as the number of spins are increased. Also note that to get access to the free energy at each temperature we simply reweight a single ISST trajectory (4.28), effectively creating $M = 25$ copies of the histogram, each representing the free energy at that temperature.

Improvement Over Standard Simulated Tempering

In this section we briefly illustrate the improvement of ISST over ST. To get an accurate comparison we implemented the ST algorithm of Nguyen *et al.* [64] with adaptive weight learning. As this method determines the weights on the fly, it is only left for us to determine the switch frequency, switch strategy and temperature distribution. We set the switching frequency to every timestep and only allow for switches between consecutive temperatures either up or down. We also distribute the temperatures linearly in β between $\beta_{\min} = 1$ and $\beta_{\max} = 3$.

For both methods we use 25 temperatures and we record the average magnetization (4.69) of a Curie Weiss system in 25 individual histograms by recording samples from a single temperature in the case of ST. In the case of ISST we reweight a single trajectory. As can be seen in Figure 4.6 this results in much better sampling for ISST than for ST. This is because ISST uses the full trajectory to compute expectations at every temperature, whereas ST only uses the pieces of the trajectory at a given temperature to compute expectations at that temperature. The procedure used in ISST reduces the statistical noise significantly for the same computational cost.

We also performed a second experiment in which we used $K = 40$ Curie Weiss spin particles. The results of recording the average magnetization m in this case are shown in Figure 4.7, where we plot the limiting result from Sec. 4.3.2 as dashed curve to provide a guide for the eye (We therefore do not expect perfect overlap of the numerical experiments and the dashed line). Again we observe the clear advantage of using the

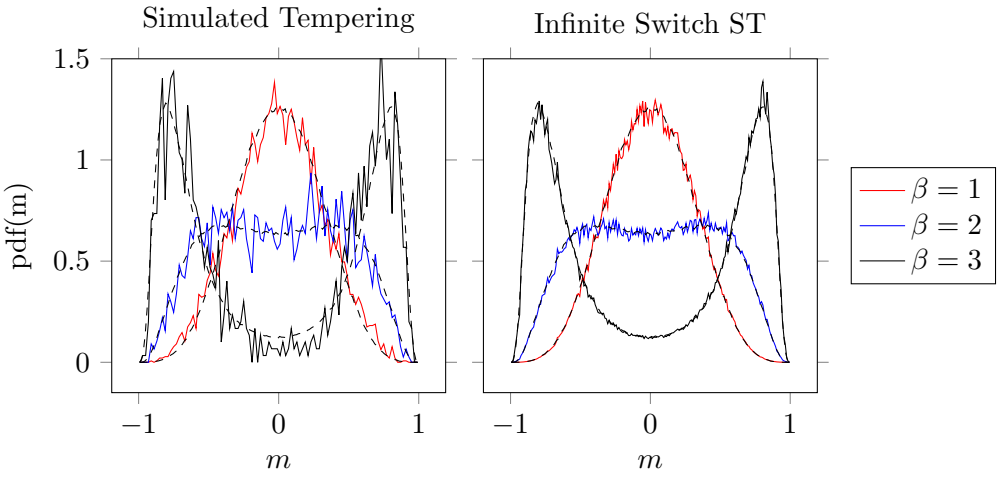

Figure 4.6: Sampling performance of the ISST algorithm compared to standard simulated tempering algorithm with adaptive weight learning proposing a temperature switch on every timestep [64]. The results are shown for a Curie Weiss system with $K = 10$ spins and external field $b = 0$, which is simple enough to be sampled by a standard Langevin scheme (with results shown as reference in dashed black). The distribution of the average magnetization (4.69) was calculated by recording a trajectory of length $N = 10^7$ in a histogram. The ISST trajectory was recorded as weighted histograms from one trajectory and the ST trajectory was recorded in the histogram corresponding to the current temperature. 25 temperatures were used, distributed linearly in β for ST and as the Legendre roots for ISST.

reweighting scheme in ISST to record the statistics.

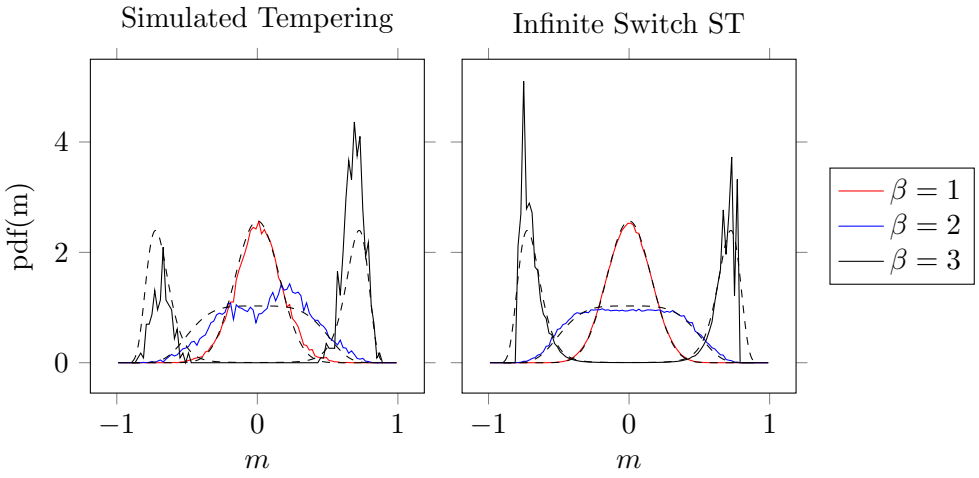

Figure 4.7: Difference in sampling performance between ST and ISST. These results are for $K = 40$ Curie Weiss spins and external field $b = 0$. The reference solution in dashed is the $K \to \infty$ limit result from Sec. 4.3.2.

We therefore conclude that using ISST significantly improves the sampling performance without introducing algorithmic complications for the same computational cost. ISST also removes the need for the practitioner to make any choices for parameters other than the limits of the desired temperature range.

Convergence of the Temperature Weights

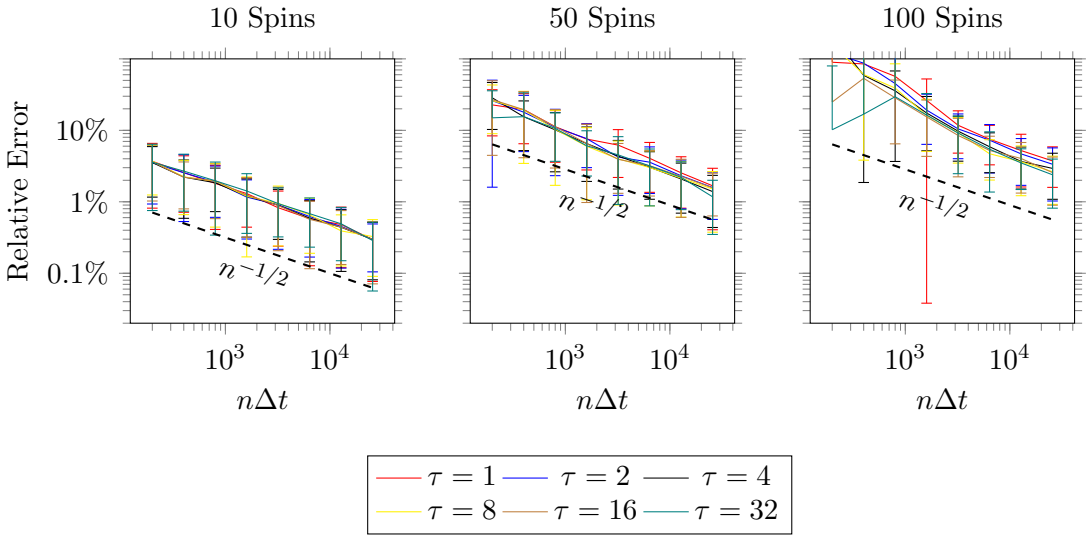

Figure 4.8: Convergence of the weight estimation using weight learning for a range of different τ. The relative error for n is calculated with respect to the weight estimate at $2n$ as given in (4.86), i.e. both the quantities were calculated as an average of 128 independent trajectories of length n and $2n$ respectively. The relative error shown in the figures, and its standard deviation (error bars), was found by averaging over 128 independent relative error estimates.

Finally we recompute the experiment of Sec. 4.3.1, confirming the conclusion that τ does not play a major role in the convergence of the temperature weights. As the long-time asymptotic weights cannot be expressed explicitly we modify the relative error such that,

$$\text{Rel.Error} = \sum_{M>i\geq 0} \frac{|\mathbb{E}_{128}\left[\omega_{i,n}\right] - \mathbb{E}_{128}\left[\omega_{i,2n}\right]|}{|\mathbb{E}_{128}\left[\omega_{i,2n}\right]|}. \tag{4.86}$$

Here, we use the notation $\mathbb{E}_{128}$ to represent an average over 128 independent ISST trajectories. We thus define the relative error as the relative difference between an average of 128 simulations of length n and an average of 128 independent ISST trajectories of length $2n$. This process is repeated 128 times to produce the points in Figure 4.8, which also shows the standard deviation of these repeated experiments as error bars.

Again, we conclude that the fixed point of the learning scheme introduced in Sec. 4.1.6 is stable for modest values of τ. We also observe that the $n^{-1/2}$ decay of the Monte Carlo sampling error dominates the accuracy of the adjustment scheme through approximation of the partition functions (4.53).

4.3.3 Alanine-12

We implemented the ISST algorithm with adaptive weight learning in the MIST package [33] and simulated the Alanine-12 molecule in vacuum using GROMACS as host code (note: the ISST algorithm is therefore also available to use with Amber 14, NAMD-Lite and LAMMPS). We used 20 temperatures at the Legendre-basis in β between 300-500K and ran the simulation for 2.2μs with a 2fs timestep. The implementation also records all the 20 individual observable weights (4.28), such that the statistics can be calculated

at any desired temperature by reweighting.

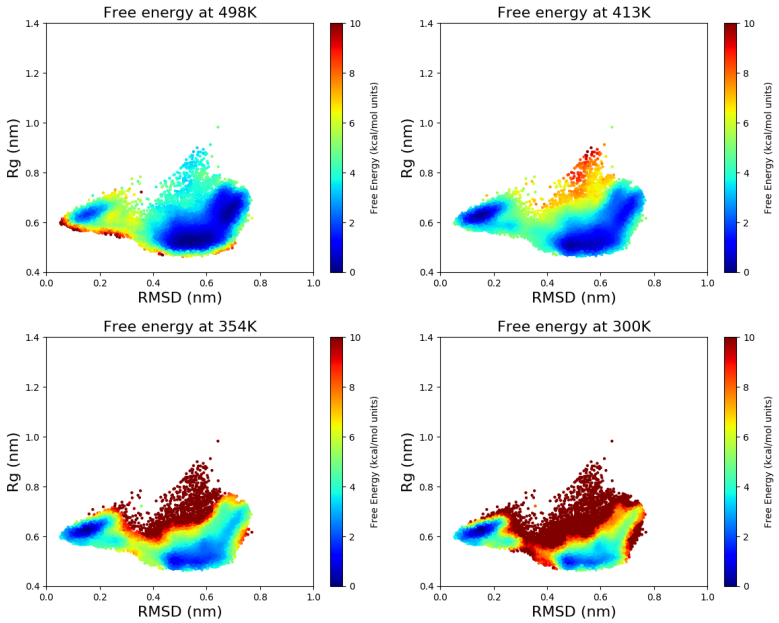

Figure 4.9: Free-energy obtained from an ISST trajectory for the in vacuum Alanine-12 to be compared directly with Figure 3.1. The simulation started from the helical configuration and ran for $2.2\mu s$ using a 2 fs timestep. The Amber96 force- field was used with a 20Å cut-off for electrostatics, constraining all bonds using the SNIP method [86].

We used GROMACS to extract the trajectory of the root mean square deviation (RMSD) and radius of gyration R_g from the initial state. In Figure 4.9 we plot the free energy of the RMSD and R_g at four temperatures. At the high temperature (top left panel) we observe 3 distinct states separated by energetic barriers. As the temperature decreases towards (from top left to bottom right) the free energy landscape changes into two distinct states separated by a large energetic barrier. As observed in [33], initializing a vanilla MD simulation in either basin at 300K will result in a skewed free energy landscape with an exponentially small probability of observing transitions between these two states.

By contrast, when using the ISST method we obtain improved sampling at all temperatures in the 300–500K range, both on the barriers and in the basins. We resolve, in detail, both the shape and the configurational states close to the states with minimal energy, giving a good overview of the structure of the free energy landscape and the available configurations at each temperature. Note that another benefit of ISST is that it significantly reduces the noise in the sampling compared to ST (this is clear from Sec. 4.3.2); ISST hence requires shorter trajectories and therefore less computational cost to achieve satisfactory results.

4.3.4 Accelerated Congressional District Sampling

In section 1.3.2 we introduced a statistical model used to sample the space of congressional district maps for North Carolina, US. This problem suffers from metastability and we applied the ISST scheme to this model with the aim of improving the collection of maps. We were ultimately not successful in this endeavour.

We suspect that this failure is the result of using temperature as the tempering variable, which maximizes the entropy for any given temperature level. Because of the energy used this meant that there was a sharp transition in temperature where the system effectively melted quickly. To obtain good performance in this case and for any similar sampling method it is not desirable to melt the system as the entropy explodes. This suggests that an investigation into the modelling aspects of the statistical model is required to avoid this phenomena.

Ultimately, this suggests that the problem of applying accelerating sampling in a problem like the congressional district allocaton has two potential solutions: (i) to reinterpret the energy to give the system a calmer response to changes in temperature or (ii) apply an entirely different method. We stress again that the temperature does not have a physical meaning for this model and the issues associated with the use of these statistical mechanical concepts lead to the numerical investigation presented in the final section.

Chapter 5

Isobaric–Isothermal Sampling Using Langevin Dynamics

Barostats provide the foundation for constant temperature and pressure simulations in molecular dynamics (MD). The scientific benefit generated from the development of efficient barostat methods is highly significant; recent examples of their use include fundamental studies of the crystal structure of water [87], the prediction of properties of materials [88], percolation studies in complex liquids [89] and molecular dynamics studies of disordered proteins [90]. This importance is further demonstrated by the active use and development of barostat implementations within various scientific MD packages, e.g. LAMMPS [91], NAMD [92] and GROMACS [93].

The goal in a NPT simulation is to provide samples from the isothermal-isobaric probability measure of a system evolving in a dynamical simulation cell, in such a way that the temperature and pressure are preserved. The statistical properties of such a system are defined by the associated isothermal-isobaric (NPT) probability distribution (1.26) with density

$$\rho_\beta(\boldsymbol{q}, \boldsymbol{p}, L) = Z^{-1} \exp\left[-\beta \left(H\left(\boldsymbol{q}, \boldsymbol{p}\right) + \mathcal{P}L\right)\right]. \tag{5.1}$$

Here $H(q, p)$ is the Hamiltonian of the N-particle system with position vector $\boldsymbol{q} \in [0, L]^{3N}$, momentum vector $\boldsymbol{p} \in \mathbb{R}^{3N}$ and $\beta = (k_B T)^{-1}$ is the reciprocal of the temperature scaled by Boltzmann's constant, $\mathcal{P}$ is the target pressure, and V is the volume of the fluctuating cell (In general, this cell may be an arbitrary parallelepiped, but it is simplest to initially present the theory in the context of a cubic simulation cell $[0, L]^3$, where L (allowed to fluctuate) is the dimension of the cube in which case $V = L^3$.). Dynamical equations can be constructed in various ways to preserve this target distribution.

Due to their relative complexity compared to constant energy or temperature methods, barostats introduce a number of choices in the design of the equations of motion, in the parameterization of those equations, and in the discretization of the resulting system. Many different techniques have consequently been proposed [94, 95, 96] typically building on the foundation proposed in the original paper of Andersen [97] which relies on an extension of phase-space and also the incorporation of auxiliary Nosé-Hoover variables as in Martyna et. al. [98].

A recent example of a constant temperature and pressure method using extended variables is the COMPEL algorithm [96], which has been applied to study chromophores [99]. This algorithm is based on auxiliary variables controlled by negative feedback loops, but it combines these with additional stochastic terms, yielding a "gentle"

stochastic method which the authors suggest provides a more accurate dynamical approximation (e.g. autocorrelation functions) than alternatives. However, despite the stochastic nature of the method, due to the feedback loop, it suffers from "ringing"–a behaviour manifested in the rapid oscillation of the volume as it approaches equilibrium. We emphasize that this effect is not a physical phenomenon but an artificial one that is due to the design of the equations used for simulation. The ringing prolongs the relaxation time of the simulation and significantly increases the computational effort needed to obtain usable results. This effect was explicitly mentioned in Goncalves et. al. [100], which illustrates the detrimental effect these numerical artifacts can have on real numerical experiments. The trade-off between dynamical fidelity and rapid convergence to thermodynamic equilibrium has been studied in [101].

5.1 A stochastic barostat based on Langevin dynamics

In recent years Langevin methods have supplanted other types of thermostats in NVT simulations of molecular systems, where they are favored due to their robustness and potential for high accuracy in statistical calculations [102]. The improvements obtained by these methods for sampling of the canonical ensemble have not gone unnoticed and there have been several attempts to create a similar barostat scheme. However, due to the variable cell dynamics this translation is nontrivial and the existing treatments rely on ad hoc assumptions which may translate into uncontrolled statistical bias or instability – limiting their accuracy, stability and reliability.

The use of a Langevin dynamics-based volume control was first introduced in the form of the *Langevin piston method* by Feller et. al [94]. This method interprets the volume control as an effective three dimensional piston with a fictitious mass that controls its fluctuations, tuned such that the fluctuations are on the same order as a sound wave travelling through the simulation cell. The method only considers deterministic particle dynamics in a cubic simulation cell and suggests an ad hoc coupling to a particle thermostat and the flexible simulation cell. The method precedes many developments in stochastic discretization obtained for constant temperature (NVT) dynamics (see below), and despite still being actively used [103] it has not been shown to be ergodic.

A more recent Langevin-NPT thermostat was developed by Gao et. al [104]. However, we are unable to reconstruct the BAOAB scheme as a logical consequence of Equation (1) of this paper, and remain unconvinced as to its mathematical validity (see below for another criticism of this article). The scheme presented in [105] appears to capture the high configurational sampling accuracy accessible using Langevin splitting. However, the scheme presented there is based on ad hoc discretization and does not fit within the framework of stochastic Lie-Trotter splitting. Another recent treatment that claims to provide a Langevin based barostat is that of Cajahuaringa and Antonelli [106], we also raise some concerns about the foundation of their method and their analysis.

Our aim is to provide a transparent, splitting-based approach to the discretization of Langevin dynamics. The partition function which appears as the normalization constant Z in (5.1) ensures that the density integrates to unity:

$$Z = \int_{[0,\infty)} \mathrm{d}L \int_{L^{3N}} \mathrm{d}\boldsymbol{q} \int_{\mathbb{R}^{3N}} \mathrm{d}\boldsymbol{p}\ \exp\left[-\beta\left(H(\boldsymbol{q},\boldsymbol{p}) + \mathcal{P}L^3\right)\right]. \tag{5.2}$$

This equation exposes the coupling between the volume L^3 and the position coordinates clearly. The complication in NPT simulation arises from this interdependence and

makes it both non-trivial to calculate the partition function (5.2) and *crucially* to sample the distribution (5.1). As first noted by Andersen [97], problems appearing as a result of the dynamical simulation cell are avoided by mapping each particle to its equivalent non-dimensional position in a cubic unit-cell. The goal is then to consider the equations of motion in this new environment. One such map is a simple rescaling of the positions such that,

$$\boldsymbol{q}_i = L\tilde{\boldsymbol{q}}_i \qquad \text{with} \qquad \tilde{\boldsymbol{q}}_i \in [0,1]^3, \quad \forall i \leq N. \tag{5.3}$$

Under this map, the isobaric-isothermal partition function in the new set of coordinates $(\tilde{\boldsymbol{q}}, \boldsymbol{p}, L)$ is,

$$\hat{Z} = \int_{[0,\infty)} \mathrm{d}L \int_{[0,1]^{3N}} \mathrm{d}\tilde{\boldsymbol{q}} \int_{\mathbb{R}^{3N}} \mathrm{d}\boldsymbol{p} \; \exp\left[-\beta\left(H(L\tilde{\boldsymbol{q}}, \boldsymbol{p}) + \mathcal{P}L^3 - \beta^{-1}N\log L^3\right)\right]. \tag{5.4}$$

The Jacobian of the mapping enters the equations of motion as a modification of the effective potential energy. It is clear to see that $\hat{Z}$ can be calculated numerically. As a consequence of this rescaling the isothermal-isobaric distribution can be sampled using the non-dimensional coordinates $(\tilde{\boldsymbol{q}}, \boldsymbol{p}, L)$.

Below we make it clear that further simplifications and desired properties can be obtained by considering a symplectic (volume preserving) map of the physical coordinates. This can be achieved by a simultaneous rescaling the momentum, i.e. $\boldsymbol{p} = L^{-1}\tilde{\boldsymbol{p}}$. In this case the potential energy is not altered as the coordinate mapping is symplectic and the Jacobian is unitary. We reiterate again that the important distinction between the original coordinates, $(\boldsymbol{q}, \boldsymbol{p}, L)$, and the rescaled coordinates $(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{p}}, L)$ is that the latter can easily be sampled and discretised.

The invariant distribution of an ergodic Langevin barostat is (5.1). The rescaling of the particle coordinates results in a system of equations with an effective position dependent particle mass matrix (variable metric), intrinsically eliminating the possibility of a direct derivation of a pure equivalent to the BAOAB method in the NPT setting (and further contradicting the claims made by Gao et al. [104]). In fact, this makes the equations of motion closer in nature to a geodesic integrator with holonomic constraints [107]. Let us state the two requirements that the correct design of a pure Langevin-barostat must respect:

1. The mapping of *both* physical variables into their non-dimensional counterparts $(\boldsymbol{q}, \boldsymbol{p}) \to (\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{p}})$ must be symplectic. This results in a scheme which preserves the volume of integrals in phase-space, which is a property important for the long finite time limit accuracy and high order of discretisations of SDEs [17, 19].

2. The discretization of the full NPT dynamics must ensure that the discretization of the effective Hamiltonian is symplectic.

The class of Langevin-based NPT methods derived in this chapter are based on the careful evaluation of both these points and are presented at the end of Section 5.2. In the following Section 5.3 the numerical properties of these schemes are presented; additionally, we demonstrate the differences between the schemes' performances by considering the characteristics of each method's auto-correlation function.

5.1.1 The Periodic Flexible Simulation Cell

The idea of the fully flexible periodic simulation cell for NPT sampling in MD derives from Parrinello and Rahman [108] and is a generalisation of the work by Andersen

[26] in the same year. These papers were so influential that a cubic periodic cell is now referred to as an *Andersen cell* and a fully flexible periodic simulation cell as a *Parrinello-Rahman cell*. In fact, as noted by Parrinello and Rahman, it is useful to denote the periodic simulation cell as a matrix $S \in \mathbb{R}^{3\times 3}$ where $S = [\boldsymbol{\ell}_x, \boldsymbol{\ell}_y, \boldsymbol{\ell}_z]$. In this case $\boldsymbol{\ell}_i \in \mathbb{R}^3$ is the cell side and the volume is the determinant of matrix S, $V = \det S$. The Andersen simulation cell with volume $V = L^3$ is in this context introduced as the diagonal matrix $S = \operatorname{diag}(L, L, L)$. The Parinello-Rahman simulation cell is on the other hand completely unconstrained and takes the form,

$$S = \begin{pmatrix} \ell_{x,0} & \ell_{y,0} & \ell_{z,0} \\ \ell_{x,1} & \ell_{y,1} & \ell_{z,1} \\ \ell_{x,2} & \ell_{z,2} & \ell_{z,2} \end{pmatrix}. \tag{5.5}$$

This definition allows the cell to rotate and translate freely. Three degrees of freedom are removed from the simulation cell if we fix it in space by restricting its allowed movements. We represent it instead by a (upper/lower) triangular matrix e.g,

$$S = \begin{pmatrix} \ell_{x,0} & \ell_{y,0} & \ell_{z,0} \\ 0 & \ell_{y,1} & \ell_{z,1} \\ 0 & 0 & \ell_{z,2} \end{pmatrix}. \tag{5.6}$$

This is the cell which we consider in the remainder of the chapter. Regardless of the shapes allowed for the simulation cell, it is common to consider periodic cells as they imitate bulk conditions. These conditions are reinforced by the *minimum image convention*, which formalises how distances between particles are calculated under periodic boundary conditions. To this end, consider a particle i at position $S^{-1}\boldsymbol{q}_i \in [0,1]^3$. This particle has a replica (copy of itself or ghost particle) at position,

$$\boldsymbol{\xi}_i = \boldsymbol{q}_i + nS\hat{\boldsymbol{q}}_i, \tag{5.7}$$

where $n \in \mathbb{Z}$ and $\hat{\boldsymbol{q}}_i$ is the unit vector of $\boldsymbol{q}_i$. The distance between any two particles in the volume bounded by S, say at position $\boldsymbol{q}_i$ and $\boldsymbol{q}_j$, is defined as,

$$\boldsymbol{r}_{ij} = \min_n \left| \boldsymbol{q}_i - \boldsymbol{q}_j + nS\hat{\boldsymbol{r}}_{ij} \right|. \tag{5.8}$$

Equation (5.8) is known as the minimum image convention.

5.2 Splitting Schemes for NPT Dynamics

Assume that there are N particles in a periodic simulation cell defined by a parallelepiped represented by the action of a 3×3 matrix S e.g (5.6). Let the position of the particles be

$$\left\{ \boldsymbol{q}_i \mid S^{-1}\boldsymbol{q}_i \in [0,1]^3 \right\}_{0<i\leq N}, \tag{5.9}$$

with momentum vectors, $\boldsymbol{p} \in \mathbb{R}^{3N}$. Represent the full set of coordinates as,

$$\boldsymbol{q} \quad = (\boldsymbol{q}_1, \boldsymbol{q}_2, \dots \boldsymbol{q}_N)^T \quad \text{with} \quad \boldsymbol{q}_i = (q_{i,x}, q_{i,y}, q_{i,z})^T,$$

with a similar definition of $\boldsymbol{p}$. Denote the mass matrix by $M \in \mathbb{R}^{3N} \times \mathbb{R}^{3N}$, typically assumed to be a diagonal matrix with positive entries $(m_1, m_1, m_1, m_2, m_2, m_2, \dots, m_N, m_N, m_N)$.

We denote by S_p the conjugate momentum to S, and let the (artificial) mass matrix

of the cell be denoted as $\mu \in \mathbb{R}^{3\times 3}$. These two variables will be related dynamically by the equation

$$\dot{S} = \mu^{-1} S_p.$$

Define also the repeated matrix $\mathcal{S} = \mathrm{diag}_N(S)$ as the $3N \times 3N$ matrix with S repeated on the diagonal, similarly also define $\mathcal{S}_p = \mathrm{diag}_N(S_p)$. With these notations, the molecular system can be described by an extended effective Hamiltonian:

$$\begin{aligned} H_{\text{eff}}\left(q, p, S, S_p\right) &= \frac{1}{2} p^T M^{-1} p + \frac{1}{2} S_p^T \mu^{-1} S_p + U(q) + \mathcal{P} \det\left(S\right), \\ &= \mathcal{K}(p, S_p) + \mathcal{U}(q, S). \end{aligned} \tag{5.10}$$

Here, $\mathcal{K}(p, S_p)$ denotes the total effective kinetic energy of the system and $\mathcal{U}(q, S)$ is its potential. The objective of an NPT barostat method is to sample the density,

$$\rho_\beta(q, p, S, S_p) = Z^{-1} \exp\left[-\beta H_{\text{eff}}\left(q, p, S, S_p\right)\right]. \tag{5.11}$$

where Z is a normalization constant. There are several choices of dynamics that are ergodic with respect to (5.11) but the treatment presented here is based on Langevin dynamics. This scheme will be based on the combination of a Lie-Trotter splitting of the Hamiltonian with judicious introduction of stochastic terms that ensure the ergodicity of the resulting method. Care must be taken in the treatment of the separate kinetic and potential terms of the total energy.

To simulate this system at constant target pressure $\mathcal{P}$, S, which defines the simulation cell, is treated as an auxiliary variable in an extended system. This introduces an additional complexity in the above system as now the physical parameters $\boldsymbol{q}$ and $\boldsymbol{p}$ are their domain depend on S. To remove this dependence it is common [97] to introduce the non-dimensional position relative to the unit cell defined by $\tilde{\boldsymbol{q}} \in [0, 1]^{3N}$ and also its corresponding momentum $\tilde{\boldsymbol{p}} \in \mathbb{R}^{3N}$. It is also natural for this mapping to be taken to be symplectic, since then the volume element in phase-space integrals is preserved. A simple choice that respects this property is the symplectic map $T(S)$,

$$\begin{pmatrix} \boldsymbol{q} \\ \boldsymbol{p} \end{pmatrix} = T(S) \begin{pmatrix} \tilde{\boldsymbol{q}} \\ \tilde{\boldsymbol{p}} \end{pmatrix}, \qquad \text{with} \qquad T(S) = \begin{pmatrix} \mathcal{S} & 0 \\ 0 & \mathcal{S}^{-T} \end{pmatrix}. \tag{5.12}$$

This map preserves the product $\mathrm{d}\boldsymbol{q} \wedge \mathrm{d}\boldsymbol{p}$ (here we denote $S^{-T} = \left[S^{-1}\right]^T$) and thus conserves phase-space volume. Introducing this mapping in the effective Hamiltonian (5.10) yields,

$$H_{\text{eff}}\left(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{p}}, S, S_p\right) = \mathcal{K}\left(\mathcal{S}^{-T} \tilde{\boldsymbol{p}}, S_p\right) + \mathcal{U}(\mathcal{S}\tilde{\boldsymbol{q}}, S). \tag{5.13}$$

It should be noted that the effective mass of the particles in the $(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{p}})$ coordinates is no longer constant: it is now dependent on the simulation cell S. This implies that the standard "BAOAB" discretization [16] (which has been shown to give high configurational sampling accuracy) cannot be used directly. Instead these equations are by their nature more similar to those of holonomically constrained Langevin dynamics [107].

In contrast to rescaling both physical variables simultaneously, a position dependent mass could be avoided by only rescaling position. This introduces the Jacobian of the map $\boldsymbol{q} \to \tilde{\boldsymbol{q}}$ in the effective potential, as explained in the introduction, to account for the coordinate rescaling,

$$\hat{\mathcal{U}}(\tilde{\boldsymbol{q}}, S) = \mathcal{U}\left(\tilde{\boldsymbol{q}}, S\right) - N\beta^{-1} \log\left(\det S\right). \tag{5.14}$$

Such a mechanism could instead be adopted here, but would not result in a symplectic splitting of the deterministic second-order dynamics. Further, note that this non-symplectic mapping also introduces a force component that scales with the number of particles and cell volume, which could have a potential destabilizing effect on the resulting dynamics by generating large forces.

From our experience in developing constant temperature schemes, incorporating a symplectic splitting of the Hamiltonian within a stochastic scheme, as in the Geometric Langevin Algorithm [109] typically generates very effective ergodic discretization methods. Furthermore, as we will show below, the effective position dependent mass matrix is a formal complication which vanishes in the final form of the discrete equations.

To sample (5.11) we introduce the extended Langevin dynamics in the non-dimensional coordinates $\tilde{\boldsymbol{q}}$ and $\tilde{\boldsymbol{p}}$,

$$\begin{aligned}
\mathrm{d}\tilde{\boldsymbol{q}} &= \nabla_{\tilde{\boldsymbol{p}}}\mathcal{K}\left(\mathcal{S}^{-T}\tilde{\boldsymbol{p}}, S_p\right), \\
\mathrm{d}S &= \nabla_{S_p}\mathcal{K}\left(\mathcal{S}^{-T}\tilde{\boldsymbol{p}}, S_p\right), \\
\mathrm{d}\tilde{\boldsymbol{p}} &= -\nabla_{\tilde{\boldsymbol{q}}}\mathcal{U}(\mathcal{S}\tilde{\boldsymbol{q}}, S) - \nabla_{\tilde{\boldsymbol{p}}}\mathcal{K}\left(\mathcal{S}^{-T}\tilde{\boldsymbol{p}}, S_p\right) + \sigma_{\boldsymbol{p}}\,\mathrm{d}\boldsymbol{W}_{\boldsymbol{p}}, \\
\mathrm{d}S_p &= -\nabla_S\mathcal{K}\left(\mathcal{S}^{-T}\tilde{\boldsymbol{p}}, S_p\right) - \nabla_S\mathcal{U}(\mathcal{S}\tilde{\boldsymbol{q}}, S) - \nabla_{S_p}\mathcal{K}\left(\mathcal{S}^{-T}\tilde{\boldsymbol{p}}, S_p\right) + \sigma_{S_p}\,\mathrm{d}W_{S_p},
\end{aligned} \tag{5.15}$$

here $\sigma_{\boldsymbol{p}}^T\sigma_{\boldsymbol{p}} = 2\gamma\beta^{-1}$ and $\sigma_{S_p}^T\sigma_{S_p} = 2\nu\beta^{-1}$. The Langevin frictions γ and ν represent the particle friction and cell friction respectively. By writing the equations in this form, the structure of the effective dynamics becomes clear, paving the way for symplectic splitting.

The Lie-Trotter splitting we construct for the dynamics given in (5.15) is as follows:

$$\begin{aligned}
\mathcal{A} &= \begin{cases} \mathrm{d}\tilde{\boldsymbol{q}} = \nabla_{\tilde{\boldsymbol{p}}}\mathcal{K}\left(\mathcal{S}^{-T}\tilde{\boldsymbol{p}}, S_p\right)\mathrm{d}t, \\ \mathrm{d}S = \nabla_{S_p}\mathcal{K}\left(\mathcal{S}^{-T}\tilde{\boldsymbol{p}}, S_p\right)\mathrm{d}t, \\ \mathrm{d}S_p = -\nabla_S\mathcal{K}\left(\mathcal{S}^{-T}\tilde{\boldsymbol{p}}, S_p\right)\mathrm{d}t, \end{cases} \\
\mathcal{B} &= \begin{cases} \mathrm{d}\tilde{\boldsymbol{p}} = -\nabla_{\tilde{\boldsymbol{q}}}\mathcal{U}(\mathcal{S}\tilde{\boldsymbol{q}}, S)\,\mathrm{d}t, \\ \mathrm{d}S_p = -\nabla_S\mathcal{U}(\mathcal{S}\tilde{\boldsymbol{q}}, S)\,\mathrm{d}t, \end{cases} \\
\mathcal{O} &= \begin{cases} \mathrm{d}\tilde{\boldsymbol{p}} = -\gamma\mathcal{S}^{-1}M^{-1}\mathcal{S}^{-T}\tilde{\boldsymbol{p}}\,\mathrm{d}t + \sigma_p\,\mathrm{d}W_p, \\ \mathrm{d}S_p = -\nu\mu^{-1}S_p\,\mathrm{d}t + \sigma_{S_p}\,\mathrm{d}W_{S_p}. \end{cases}
\end{aligned} \tag{5.16}$$

Implicitly, we also assume $\mathrm{d}X = 0$ for any parameter not explicitly expressed. Note also the extra equation in $\mathcal{A}$ resulting from the domain dependence in the effective kinetic energy, resulting from the Hamiltonian structure given in (5.15). It is clear from $\mathcal{A}$ and $\mathcal{B}$ that these components are based on a symplectic splitting of the rescaled Hamiltonian (5.13).

Both the $\mathcal{B}$ and $\mathcal{O}$ in (5.16) can be solved exactly whilst $\mathcal{A}$ cannot. This is in contrast with the NVT schemes in Euclidean space [16] where all the individual components including $\mathcal{A}$ can be solved explicitly, a direct consequence of the position independent mass matrix. The $\mathcal{A}$ step must therefore be handled differently in the pressure controlled system. Here we propose to split it into two partitioned steps,

$$\mathcal{A}^1 = \left\{\mathrm{d}S = \nabla_{S_p}\mathcal{K}\left(\mathcal{S}^{-T}\tilde{\boldsymbol{p}}, S_p\right)\mathrm{d}t, \qquad \mathcal{A}^2 = \begin{cases} \mathrm{d}\tilde{\boldsymbol{q}} = \nabla_{\tilde{\boldsymbol{p}}}\mathcal{K}\left(\mathcal{S}^{-T}\tilde{\boldsymbol{p}}, S_p\right)\mathrm{d}t, \\ \mathrm{d}S_p = -\nabla_S\mathcal{K}\left(\mathcal{S}^{-T}\tilde{\boldsymbol{p}}, S_p\right)\mathrm{d}t. \end{cases}\right. \tag{5.17}$$

It should be noted that both $\mathcal{A}^1$ and $\mathcal{A}^2$ can be solved exactly. However $\mathcal{A}^1$ contains

a subtle constraint, $\mathrm{d}\tilde{\boldsymbol{q}}/\,\mathrm{d}t = 0$ and $\mathrm{d}\tilde{\boldsymbol{p}}/\,\mathrm{d}t = 0$, which is not automatically satisfied. Intuitively, this constraint encodes the behaviour that when S is changed the relative position and momentum should remain constant. At the discrete level, this constraint is trivial to implement and all the individual steps can now be solved exactly.

Also note that $S^T M_i S$ is symmetric for the particle mass M_i diagonal. This implies that for every $i \in N$,

$$S^T M_i S = \left(M_i^{1/2} S\right)^T M_i^{1/2} S, \tag{5.18}$$

which is utilized to simplify $\mathcal{O}$.

By exploiting $\mathrm{d}S = 0$ in $\mathcal{A}^2, \mathcal{B}$ and $\mathcal{O}$ (5.16) and (5.17) can be solved to give,

$$\begin{aligned}
A_h^1 &= \begin{cases} \boldsymbol{q}_n = \mathcal{S}_{n+1}\mathcal{S}_n^{-1}\boldsymbol{q}_n, \\ \boldsymbol{p}_n = \mathcal{S}_{n+1}^{-T}\mathcal{S}_n^T\boldsymbol{p}_n, \\ S_{n+1} = S_n + h\mu^{-1}S_{p,n}, \end{cases} \\
A_h^2 &= \begin{cases} \boldsymbol{q}_{n+1} = \boldsymbol{q}_n + hM^{-1}\boldsymbol{p}_n, \\ S_{p,n+1} = S_{p,n} - h\nabla_S\left(\tilde{\boldsymbol{p}}_n^T\mathcal{S}^{-1}M^{-1}\mathcal{S}^{-T}\tilde{\boldsymbol{p}}_n\right)\big|_{\mathcal{S}=\mathcal{S}_n}, \end{cases} \\
B_h &= \begin{cases} \boldsymbol{p}_{n+1} = \boldsymbol{p}_n - h\nabla_q U(q), \\ S_{p,n+1} = S_{p,n} - h\left(\nabla_S U(\mathcal{S}\tilde{\boldsymbol{q}}_n) + \mathcal{P}\nabla_S \det(S)\right)\big|_{S=S_n}, \end{cases} \\
O_h &= \begin{cases} \boldsymbol{p}_{n+1} = e^{-\gamma h}\boldsymbol{p}_n + \sqrt{\beta^{-1}\left(1-e^{-2\gamma h}\right)}M^{1/2}R_n, \\ S_{p,n+1} = e^{-\nu h}S_{p,n} + \sqrt{\beta^{-1}\left(1-e^{-2\nu h}\right)}\mu^{1/2}\hat{R}_n. \end{cases}
\end{aligned} \tag{5.19}$$

Here, $R_n, \hat{R}_n \sim \mathcal{N}(0,1)$ and h is the timestep. We have also made the constraint in A^1 explicit. Because A^1 and A^2 are approximations of $\mathcal{A}$ they should be combined to form an estimator of this step. This could be achieved by a symmetric sequence of arbitrary length, combining different permutations of A^1 and A^2. We found a sequence of length three to be sufficient. In Section 5.3, methods based on the symmetric composition

$$A_h^{xyx} := A_{h/2}^x A_h^y A_{h/2}^x, \tag{5.20}$$

are investigated for $x, y \in \{1, 2\}$. We also focus our attention on schemes which are overall symmetric as it is straightforward to show that any symmetric integrator consisting of five substeps, integrated with a method of order two at least, provides second order convergence with respect to the timestep h. Note that this assumes that one can solve the A step exactly. The six schemes compared in the next section are,

$$\begin{aligned}
&B_{h/2}A_{h/2}^{121}O_hA_{h/2}^{121}B_{h/2}\,, && B_{h/2}A_{h/2}^{212}O_hA_{h/2}^{212}B_{h/2} \\
&A_{h/2}^{121}B_{h/2}O_hB_{h/2}A_{h/2}^{121}\,, && A_{h/2}^{212}B_{h/2}O_hB_{h/2}A_{h/2}^{212} \\
&O_{h/2}B_{h/2}A_h^{121}B_{h/2}O_{h/2}\,, && O_{h/2}B_{h/2}A_h^{212}B_{h/2}O_{h/2}\,,
\end{aligned} \tag{5.21}$$

which should all be second order schemes as is explained below.

5.2.1 Brief Aside on A, μ and Friction

Let us briefly discuss the implementation of the constrained $\mathcal{A}$ step and its splitting into A^1 and A^2 in more detail. As outlined above, the splitting in (5.20) is a direct consequence of the position dependence in the effective mass in (5.15) which implies that a pure BAOAB scheme does not exist for these equations. To identify any differences between the two implementations (121 or 212) the generators for these steps are studied.

To simplify the analysis, consider a cubic cell with scalar dynamic volume V and corresponding momentum V_p. The Fokker-Planck operator for $\mathcal{A}$ in this setting for the system is,

$$\mathcal{L}^\dagger_{\mathcal{A}} = V^{-2/3}\tilde{\boldsymbol{p}}^T M^{-1}\nabla_{\tilde{\boldsymbol{q}}} + \mu^{-1}V_p\partial_V - \frac{1}{3}V^{-5/3}\tilde{\boldsymbol{p}}^T M^{-1}\tilde{\boldsymbol{p}}\partial_{V_p}. \tag{5.22}$$

From this it is straightforward to write the operators of the individual steps. Following the techniques developed in the literature [13, 16] the Baker-Campbell-Hausdorff formula is employed to derive the order correction for the timestep h of (5.20) as,

$$\begin{aligned} \mathcal{O}(1) \quad & \mathcal{L}^\dagger_0 = \mathcal{L}^\dagger_{\mathcal{A}}, \\ \mathcal{O}(h^2) \quad & \mathcal{L}^\dagger_2 = \frac{1}{12}\left[\mathcal{L}^\dagger_y, \left[\mathcal{L}^\dagger_y, \mathcal{L}^\dagger_x\right]\right] - \frac{1}{24}\left[\mathcal{L}^\dagger_x, \left[\mathcal{L}^\dagger_x, \mathcal{L}^\dagger_y\right]\right], \end{aligned} \tag{5.23}$$

where $\mathcal{O}(h)$ is zero. It is of course possible to solve $\mathcal{L}^\dagger_2\rho$ to obtain the exact perturbation of the different schemes although, by inspection, it is clear that $\mathcal{L}^\dagger_2 \sim \mathcal{O}\left(\mu^{-1}\right)$. This indicates that both estimators of the form (5.20) should approximate $\mathcal{A}$ closer in the large simulation cell mass limit, with the limit $\mu \to \infty$ reducing to the exact dynamics. Additionally, it also seems to suggests that A^{121} and A^{212} are the same in this limit. This is not surprising as, intuitively, it is also the limit where the effective dynamics collapses to NVT simulation.

The use of a modestly large value for the cell mass is likely to stabilize the dynamics by introducing a time-scale separation between the particle and cell dynamics. By increasing the cell mass, its evolution slows allowing the particles to settle into equilibrium before the cell resizes. Furthermore, because the noise is injected directly in the cell momentum, an increase in the cell mass should not affect ringing.

The situation becomes more complex when considering the choice of friction parameters ν and γ. It will most likely be useful to tune these parameters in situations when a time-scale separation between the cell and particle dynamics, through the tuning of the cell mass, is not sufficient to achieve satisfactory results.

5.3 Numerical Experiments

In this section we investigate the numerical properties of the methods derived above. Note that our experiments only consider toy systems and we have not tuned the different parameters of the algorithm beyond generating stable simulations. Sensible parameters are problem dependent and should be chosen with care in physical systems, such that the volume fluctuations of the simulation cell are of the expected order.

In all the simulations below, we consider a fully flexible simulation cell where we remove all cell translational and rotational degrees of freedom. Several different cell geometries and, indeed, relations between the cell-coordinates are possible to consider, which could potentially aid simulations in different situations. We are only interested in demonstrating the properties of the algorithm and as such we consider the least restrictive conditions.

Consider a system of 512 Lennard-Jones particle cluster in $d = 3$ with the familiar pair-potential,

$$U_{LJ}(r) = 4\varepsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right). \tag{5.24}$$

Here we choose $\sigma = \varepsilon = 1$ and set $M_i = I_3$ for all particles (where I_3 represent the

unit matrix in $\mathbb{R}^3$). We also set the target pressure to $\mathcal{P} = 0.1$ and target temperature as $T = 1.25$. We implemented our own code using parallelisation and cell method to improve the performance of the integration.

As we mentioned above, we did not tune the algorithmic parameters to represent any specific system and we tried to keep the parameters close to unity where possible. We found that the simulation was stable for a long time, under mild damping with $\gamma = \nu = 1$. With these friction coefficients the simulation diffuses quickly to equilibrium, which is observed in Figure 5.1.

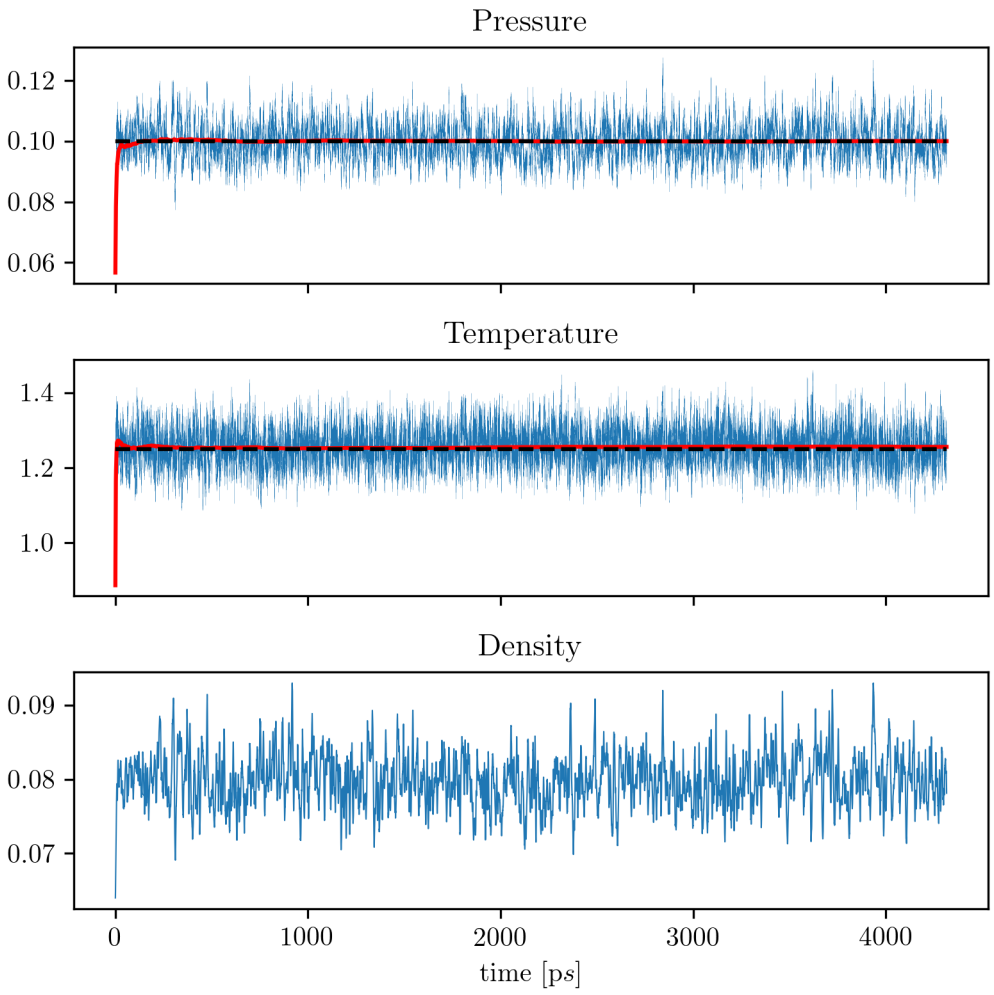

Figure 5.1: Three figures illustrating the excellent properties of the Langevin barostat $B_{h/2}A^{121}_{h/2}O_hA^{121}_{h/2}B_{h/2}$ using a $9fs$ timestep for a $4.5ns$ trajectory. Instantaneous values are shown blue, long term average in red and the expected long-term average in dashed black. The initial density was taken as $\rho_0 = 0.06$.

In some instances, most likely as a result of the LJ potential, the simulation cell would oscillate rapidly forcing particles close together and causing blow-up events. To remedy this behaviour we slowed the evolution of the simulation cell by tuning the simulation cell mass by setting $\mu = \text{diag}(10, 10, 10)$. This has the effect of slowing the time-scale evolution of the simulation cell compared to the particles. This was found to stabilise the long-run simulations.

Our main aim was also to illustrate the properties of the methods in the large timestep limit. The timestep in Figure 5.1 is $h = 0.004$. Using this timestep and the parameters of Argon to map it to physical time we find that this value corresponds to a timestep of $9f$s.

In Leimkuhler and Matthews [16] it was established that a Lie-Trotter splitting, here corresponding to $B_{h/2}A^{121}_{h/2}O_hA^{121}_{h/2}B_{h/2}$, performs well. As this scheme collapses to the BAOAB canonical sampling scheme in the large mass limit, we anticipated that this type of scheme would also perform well in the barostat setting. We also favour this scheme over $B_{h/2}A^{212}_{h/2}O_hA^{212}_{h/2}B_{h/2}$, as the latter scheme discretizes the simulation cell update with a time-step $h/2$ instead of $h/4$. This could help stabilise calculations by resolving the motion of the cell better.

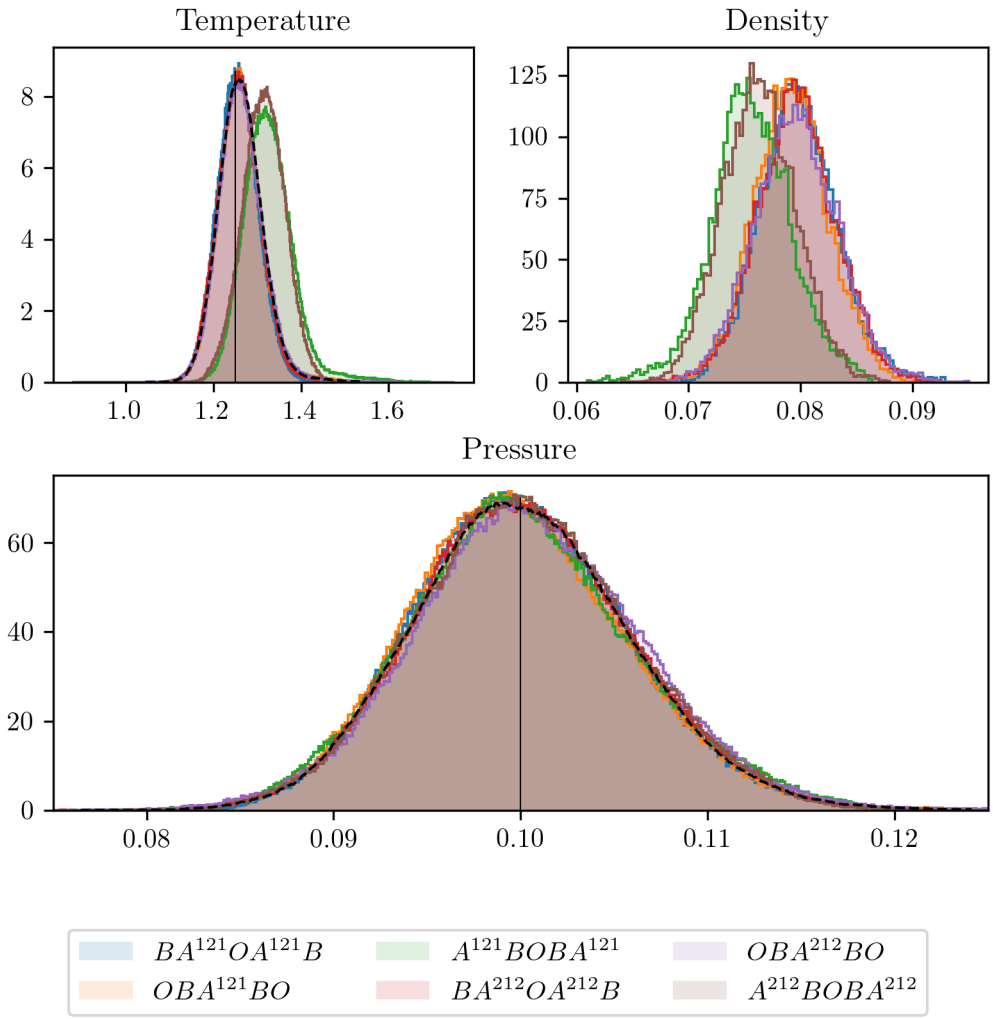

Figure 5.2: Histograms of the different observables are shown for all the methods using a trajectory of $9n$s length with a $9f$s timestep. Interestingly, all methods have an apparent similar performance in approximating the pressure, although the ABOBA methods appear to show a significant rise in the temperature and consequent lowering of the density to maintain a pressure close to the target. The target mean is shown vertically in black and a reference simulation using $\mathrm{BA}^{121}\mathrm{OA}^{121}\mathrm{B}$ with a $2f$s timestep is shown in dashed black (with an error in the mean of 0.5% in temperature and 1% in pressure).

We fixed all parameters, as described above, and performed simulations with the six different schemes (see Equation (5.21)), to investigate the difference in performance between them. The results of these experiments are shown in Figure 5.2 and clearly suggest that the two ABOBA methods are inferior. Interestingly, both methods correctly adjust the simulation cell volume as a consequence of operating at a higher temperature, which surprisingly causes the pressure to remain close to its target. This is a peculiar feature which we note is most likely a result of judicious cancellation of errors in the particular observable, which is both momentum and position dependent. On the other hand, temperature is measured using kinetic energy and shows a clear bias in the average not present in the other methods. The poor conservation of momentum temperature has been noted before for ABOBA, again for canonical sampling [102]. It is reassuring to find that the same property remains true for barostats based on a similar type of splitting.

It is more difficult to separate the performances of the BAOAB and OBABO type schemes. To distinguish these we studied the auto-correlation function for both pressure and temperature. These results are show in Figure 5.3. The graph shows that pressure

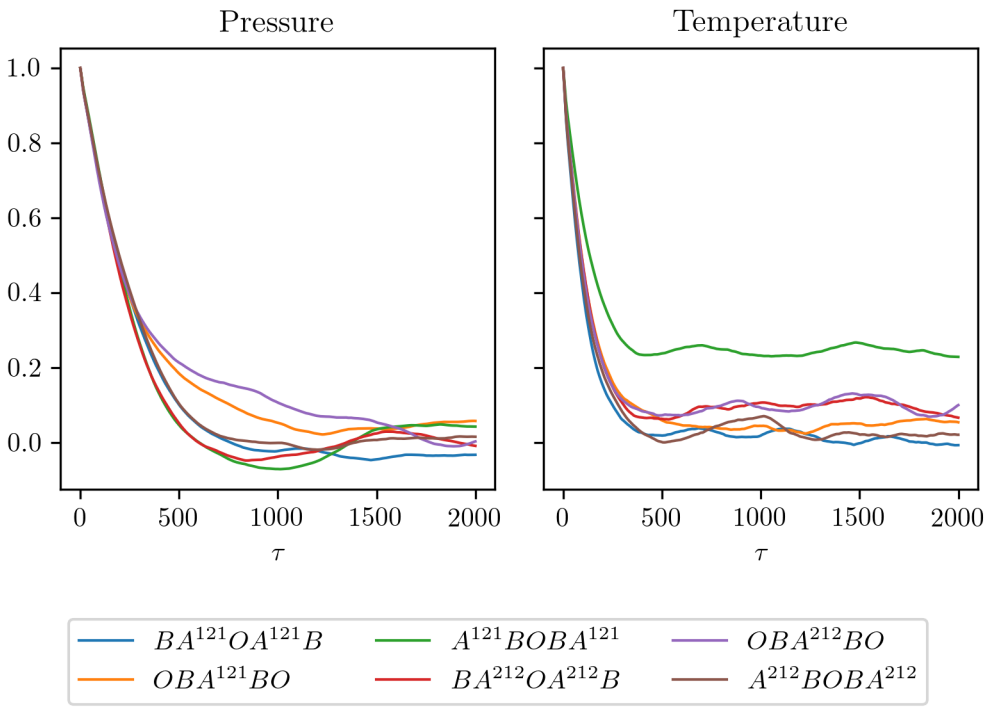

Figure 5.3: The centred auto correlation function is shown for both pressure and temperature using the six different methods for a trajectory of length $9ns$ using a $9fs$ timestep. The target pressure was taken as $\mathcal{P} = 0.1$ with temperature $T = 1.25$.

decorrelates faster for both BAOAB methods than for the OBABO methods, with an advantage for $BA^{121}OA^{121}B$ over $BA^{212}OA^{212}B$. All methods seem to display similar decorrelation in temperature with the exception of $A^{121}BOBA^{121}$ which introduces a stronger bias, seen in Figure 5.2. We attribute this to the fact that the method appeared to be close to its stability threshold and struggles to remain stable.

We also observe an advantage for the two BAOAB methods over the OBABO methods which is manifested in terms of faster convergence. This indicates that we expect both BAOAB methods to converge faster to equilibrium and as mentioned just

above, we favour $\mathrm{BA}^{121}\mathrm{OA}^{121}\mathrm{B}$ over $\mathrm{BA}^{212}\mathrm{OA}^{212}\mathrm{B}$, because of the first method's better resolution of the simulation cell dynamics.

5.3.1 Pressure Conservation Under Simulated Tempering

The simulated tempering (ST) method [110] with adaptive weight learning in Chapter 3.2 [111], is often utilized to accelerate the sampling of MD simulations. When coupled with barostats, the introduction of accelerated sampling techniques could destabilize the algorithm, complicating the parameter tuning and consequently resulting in a poor performance of the volume control.

In particular, coupling ST with a barostat could be detrimental for the conservation of pressure and relaxation of volume as the system reacts to the new temperature. This could introduce volume ringing which destabilizes the algorithm by resonating the system out of control. By tuning the simulation cell mass and friction these effects can be dampened but would impact the volume exploration. All this generates a somewhat complicated tuning procedure for all model parameters, including the switching frequency which should be as small as possible.

In this section we investigate the performance of the $B_{h/2}A^{121}_{h/2}O_hA^{121}_{h/2}B_{h/2}$ scheme when coupled with ST for the Lennard-Jones system described in the previous section. In Figure 5.4 we display the excellent performance of the scheme. It is clear from this figure that the barostat responds well to changes in temperature, even the rapid increase around the 250 ps where volume is controlled to maintain the target pressure without introducing any significant ringing. After this rapid initial change in volume and temperature, the simulation remains stable whilst the temperature is explored. In the panel labelled *Simulated Tempering* the measured instantaneous temperature trajectory is shown in blue and in red the trajectory of the target temperature as specified by the ST algorithm.

5.3.2 Merceds-Benz Potential

The Mercedes-Benz potential was first introduced by Ben-Naim [112] in 1971 as a simple $d = 2$ water model. It can be viewed as an extension of the Lennard-Jones potential with an additional anisotropic term that induces rotation,

$$\begin{aligned} U_{MB}\left(\boldsymbol{q}_i, \boldsymbol{q}_j, Q_i, Q_j\right) &= U_{LJ}(r_{ij}) \\ &\quad + \varepsilon_{HB} G\left(r_{ij} - r_{HB}\right) \\ &\quad \times \sum_{k,l=1}^{3} G\left(Q_i^T \boldsymbol{a}_{ik} \cdot \boldsymbol{u}_{ij} - 1\right) G\left(Q_j^T \boldsymbol{a}_{jl} \cdot \boldsymbol{u}_{ij} + 1\right). \end{aligned} \tag{5.25}$$

Here $r_{ij} = |\boldsymbol{q}_i - \boldsymbol{q}_j|$. Because of the anisotropic term, the notation of the model requires some explanation. We refer to the original paper for an in-depth description but note that each of the three arms, in Figure 5.5, represent the mean of a Gaussian of the form: $G(x) = \exp\left[-x^2/2\sigma_{HB}\right]$. The evolving $2d$ rotational matrix, that determines the orientation of each particle, is labeled Q_i to indicate the orientation of particle i. The unit vector along the separation between two particles labeled i and j, i.e. with magnitude r_{ij} is labeled $\boldsymbol{u}_{ij} = \boldsymbol{r}_{ij}/r_{ij}$.

To simulate the dynamics of the particles we use a rigid body Langevin scheme [113] with a rotational splitting of the BAOAB form, which exactly mimics the translational degrees of freedom. We stress that the single rotational degree of freedom is

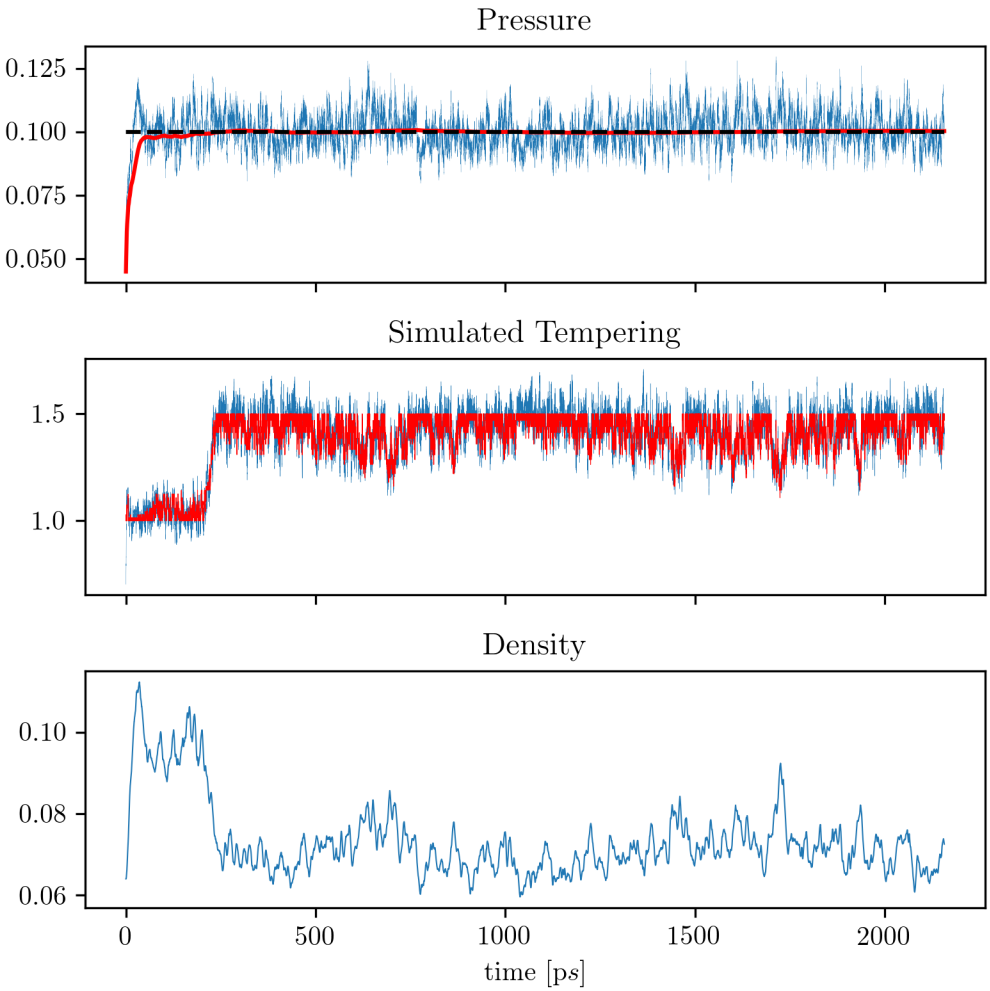

Figure 5.4: A figure that illustrates the performance of the $B_{h/2}A^{121}_{h/2}O_hA^{121}_{h/2}B_{h/2}$ using a $9fs$ timestep when coupled with ST.

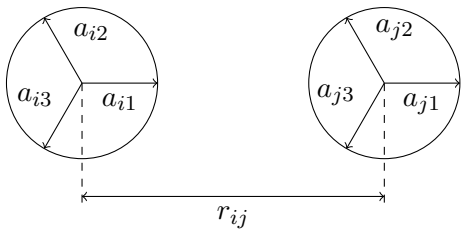

Figure 5.5: A cartoon of the Mercedes-Benz geometry is shown with the labelling of each of the three arms for two neighbourhood particles.

thermalised in these experiments. The dynamics is then straightforward to simulate with the barostat equations of motion, as given in the previous section.

Since the model contains a large number of modelling parameters, we summarise them in the following table:

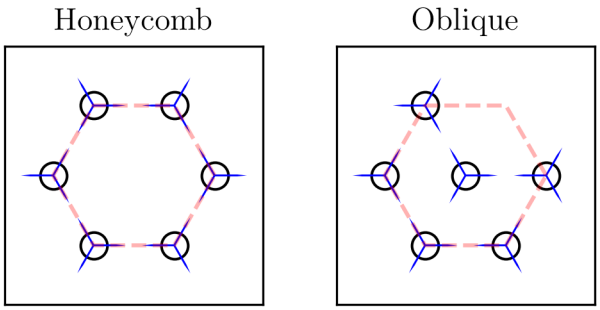

Figure 5.6: Plots of the two distinct crystal structures of the Mercedes-Benz water model are shown. Both these structures are obtained when using parameters from Table 5.3.2.

$\varepsilon_{LJ} = 0.1,$	$\varepsilon_{HB} = -1.0$
$\sigma_{LJ} = 0.7,$	$\sigma_{HB} = 0.085$
$m = 1.0,$	$I = 0.01126$
	$r_{HB} = 1.0$

Here I denotes moment of inertia. From now on we fix the simulation scheme parameters as: $\mu = 5.0$, $\nu = 2.0$, $\gamma = 1.0$ (including rotational friction) and $h = 0.01$. The $N = 1024$ particles were initialised on a honeycomb grid in a periodic 2 dimensional cell with a long range force cutoff at $r_{\text{cutoff}} = 4r_{HB}$. Note that these are the parameters used in the original paper, but also more recently in the following papers [114, 115, 116].

To the best of our knowledge, Scukins et al. [115] is the only molecular dynamics implementation of the original Mercededs-Benz potential. However, their implementation is ad hoc and uses outdated non-ergodic methods, even for the thermostat. Comparisons of our results and theirs is therefore difficult. Our intention is not to make a full comparison with either of the remaining two papers [114, 116], that used MCMC sampling schemes. We do however note that we see a significant difference in both the order and sensitivity of the crystal structures to pressure changes compared to these studies. This may be a result of the dynamical aspect of the simulations resulting from the choice of moment of inertia which is not present in those papers.

With the choice of parameters given in Table 5.3.2, the model is known to have two distinct metastable crystal structures, shown in Figure 5.6. We adopt the naming convention introduced by Sing and Bagchi [114], but in contrast to their experiments and in linewith the original Ben-Naim paper [112], we chose a value of $\varepsilon_{LJ} = 0.1$ instead of $\varepsilon_{LJ} = 0.4$. We consequently favoured the use of $\varepsilon_{LJ} = 0.1$ since σ_{HB} is chosen small enough to only allow for single hydrogen bonds to form in the first place.

A useful order parameter introduced to detect an overall change from a honeycomb dominated state to an oblique dominated one, is based on a spherical order parameter given by Auer and Frenkel [117]. This order parameter detects six-fold symmetry and is zero for the honeycomb state and one for the oblique state. Let,

$$\boldsymbol{q}_6(i) = \sqrt{\frac{4\pi}{13}\sum_{m=-6}^{6}|\boldsymbol{q}_{6m}(i)|^2}, \quad \text{where} \quad \boldsymbol{q}_{6m}(i) = \frac{1}{N_b(i)}\sum_{j=1}^{N_b(i)} Y_{6m}(\boldsymbol{u}_{ij}). \tag{5.26}$$

Here Y_{6m} is the m component spherical harmonic of the unit vector $\boldsymbol{u}_{ij}$ separating the neighbourhood particle j and the particle i. We define any particles within a distance

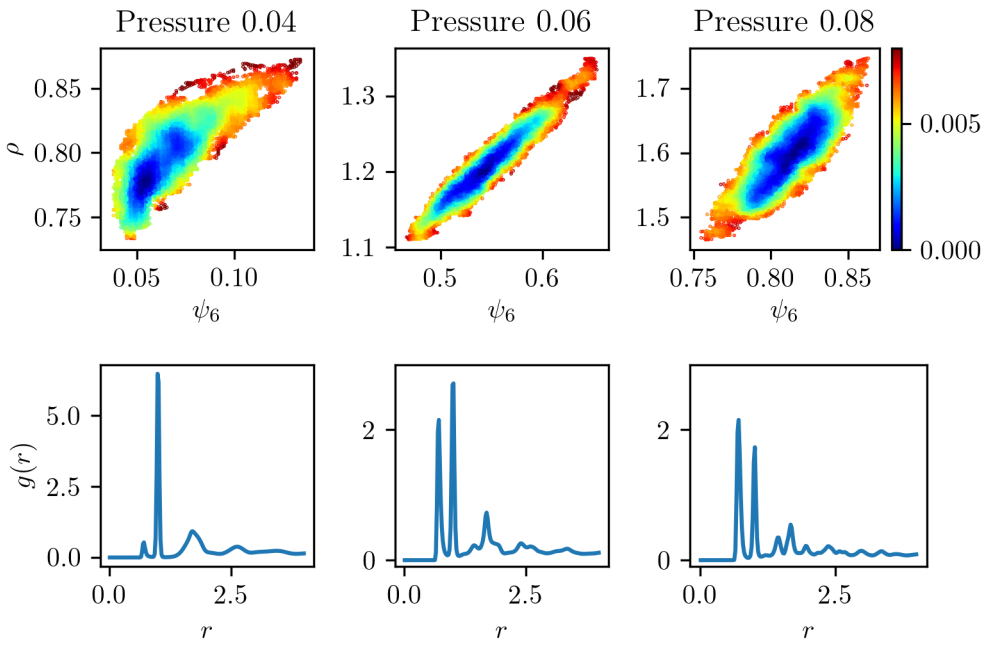

Figure 5.7: (top) Heat maps are shown for three values of the target pressures increasing from 0.04 in steps of 0.02. Blue indicates energetically favourable states. (bottom) Plots of the radial distribution function are shown for each corresponding trajectory at target pressures fixed in each column. The temperature in these experiments was fixed at $T = 0.05$ and each trajectory is of length 5×10^5 with a timestep of $h = 0.01$.

of $1.225\sigma_{LJ}$ as being neighborhood particles. From this we define the instantaneous system averaged order parameter,

$$\psi_6 = \frac{1}{N} \sum_{i}^{N} \boldsymbol{q}_6(i). \tag{5.27}$$

This treatment is similar to [114].

In Figure 5.7 we show results from simulations at target pressures increasing in steps of 0.02, starting in a honeycomb configuration at number density $\rho = N/V = 0.7$ for a simulation of length 5×10^5 with a timestep of $h = 0.01$. We exclude the first 3000 steps in which all systems reached equilibrium. In the top row of Figure 5.7 we show the heatmap of the order parameter ψ_6 given by (5.27) and the number density $\rho = N/V$, in the bottom row we plot the radial distribution functions associated with each experiment. Each column represents an increase of the target pressure in steps of 0.02 going from left to right.

Please note that the axes in the top row differ significantly between each panel to allow the full detail of each metastable state to be shown clearly. In the first column of Figure 5.7 we see evidence of the metastable honeycomb state as indicated by the low value of the order parameter and a number density between $0.7 - 0.8$. This is in line with the findings of other papers that have found the honeycomb state to be stable at these conditions, from this we connect the tallest peak at $r = 1$ of the radial distribution function $g(r)$ to the honeycomb state. This fact can also be demonstrated clearly by inspection in Figure 5.8 where we plot the final frame of each trajectory. The

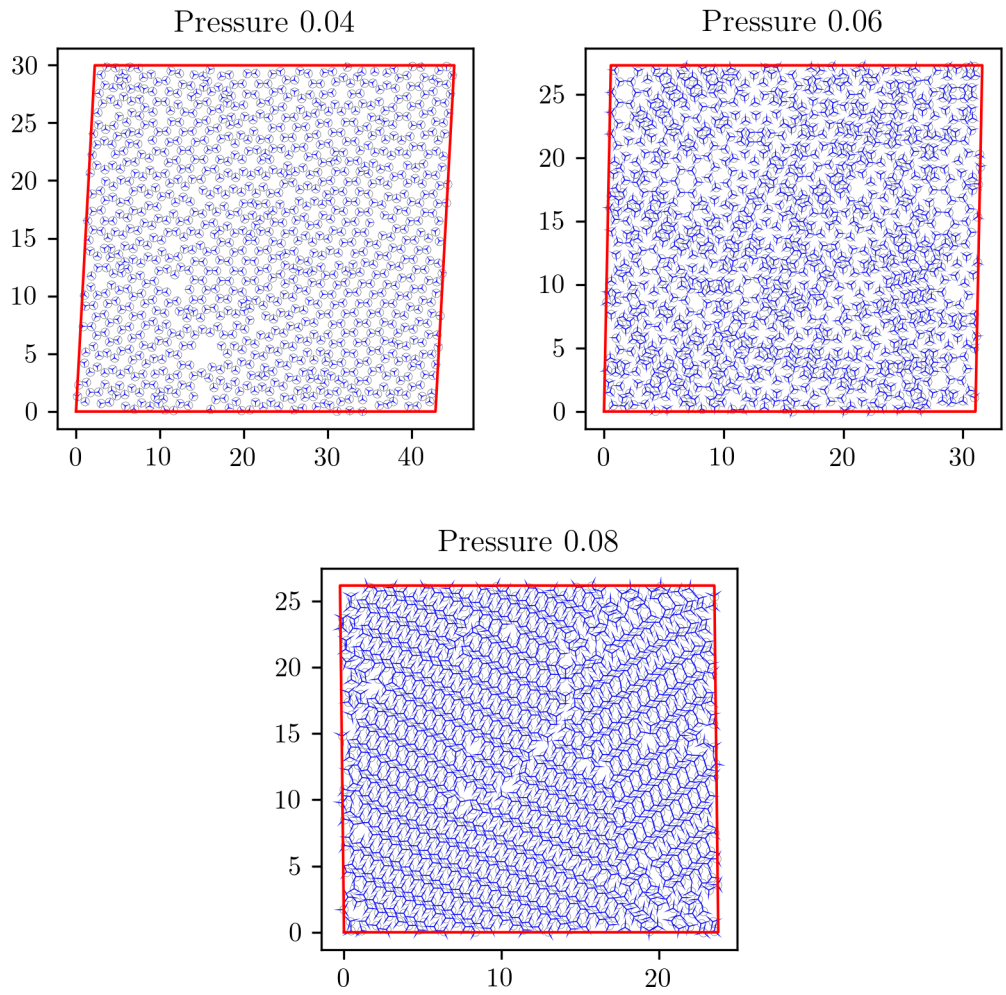

Figure 5.8: Three frames are show for the final configuration obtained for each trajectory in Figure 5.7.

frame corresponding to a target pressure of 0.04 clearly shows a global structure which is dominated by the honeycomb state.

When the system is compressed further, by setting the target pressure to 0.06, we find an elongated state which fluctuates substantially in the number density. Furthermore, a similarly large fluctuation in the order parameter with an overall value of 0.5 suggests that this is a state which is not dominated by either of the crystal structures from Figure 5.6. We elaborate that this a result of spontaneous crystallisation into a locally oblique state followed by a slow decay. A similar behaviour was noted in [114] which called such configurations of four particles "*tetratic defects*". This is further verified by inspection via Figure 5.8 in which we observe honeycomb areas interlaced by regions of higher density that resemble the oblique structure but we also see a significant number of tetratic defects.

Setting the target pressure to a value of 0.08 results in a further increase in both the order parameter and number density. Note that the shape of this well is similarly elongated as for the previous target pressure, but with a slight concentration in density. With an absolute value of the order parameter around 0.8, this is not a pure oblique

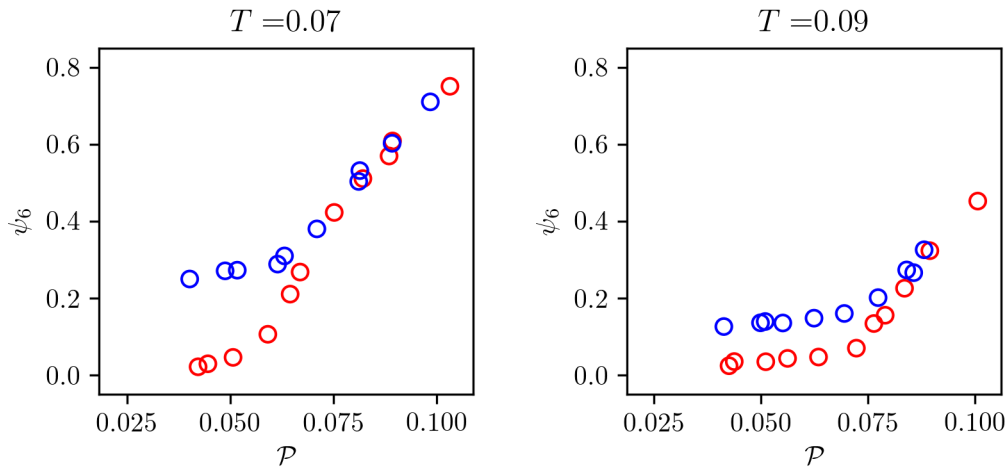

Figure 5.9: The hysteresis curves obtained for two temperatures by successive changes to the pressure. We let the system reach equilibrium in pressure and the data displayed are averages over the final 10 percent of the trajectory for each such value. We indicate the compression of the system by red and inflation in blue. The initial condition for each trajectory was taken as 0.04 in Figure 5.8.

state. This is confirmed by the final frame in Figure 5.8 which shows long strands forming separated into regions of differing alignment. A key benefit of using barostats with precise and accurate pressure control is that these types of states can be observed and studied reliably.

We also investigate the system's response to changes in pressure. These experiments involved starting with the initial configuration identical to the frame labeled 0.04 in Figure 5.8. We increase the pressure to 0.1 after which it is decreased again in the same number of step to the initial pressure of 0.04. The two branches obtained for these experiments using two target temperatures are plotted in Figure 5.9 and is in each case distinguished by red circles for compression and blue for expansion. At each of the 22 target pressures we let the system's pressure settle into equilibrium and the recorded values are averaged over the final 10% of each such sub-trajectory. We see that as the system deflates we bifurcate from the initial compression branch, emitted at a different pressure for each temperature.

In Figure 5.11 below we plot the first and final frames obtained by each trajectory for the two temperatures. From these we conclude that the bifurcation in Figure 5.9 is a result of the slow decay of the denser oblique crystal structure, which is visible in the final frames at both temperatures (especially in the lower right panel).

We also see evidence of a third crystal structure in all four panels, which we have labeled as *intermediate* in Figure 5.10. It is unclear whether this structure is an artifact of the parameter choice of value for moment of inertia. Most of the available papers use MCMC sampling and it therefore makes any direct comparison difficult. We note that none of the available papers cited in this section report on this third structure in their experiments and we do not make any claim as to whether this structure should be present at these conditions.

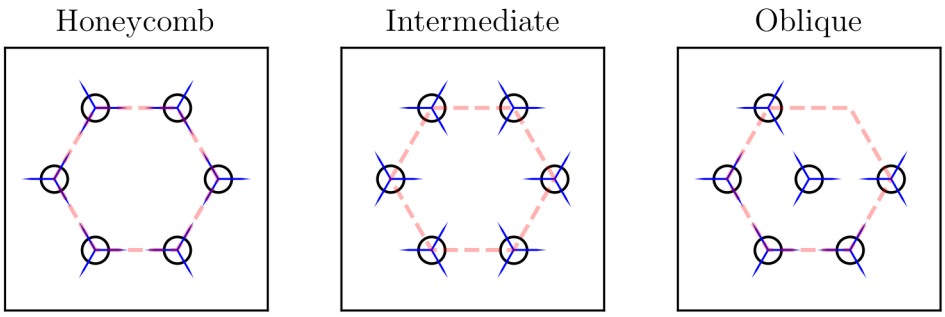

Figure 5.10: The figure shows the potential and third crystal structure labeled intermediate, which appears in all four frames of Figure 5.11.

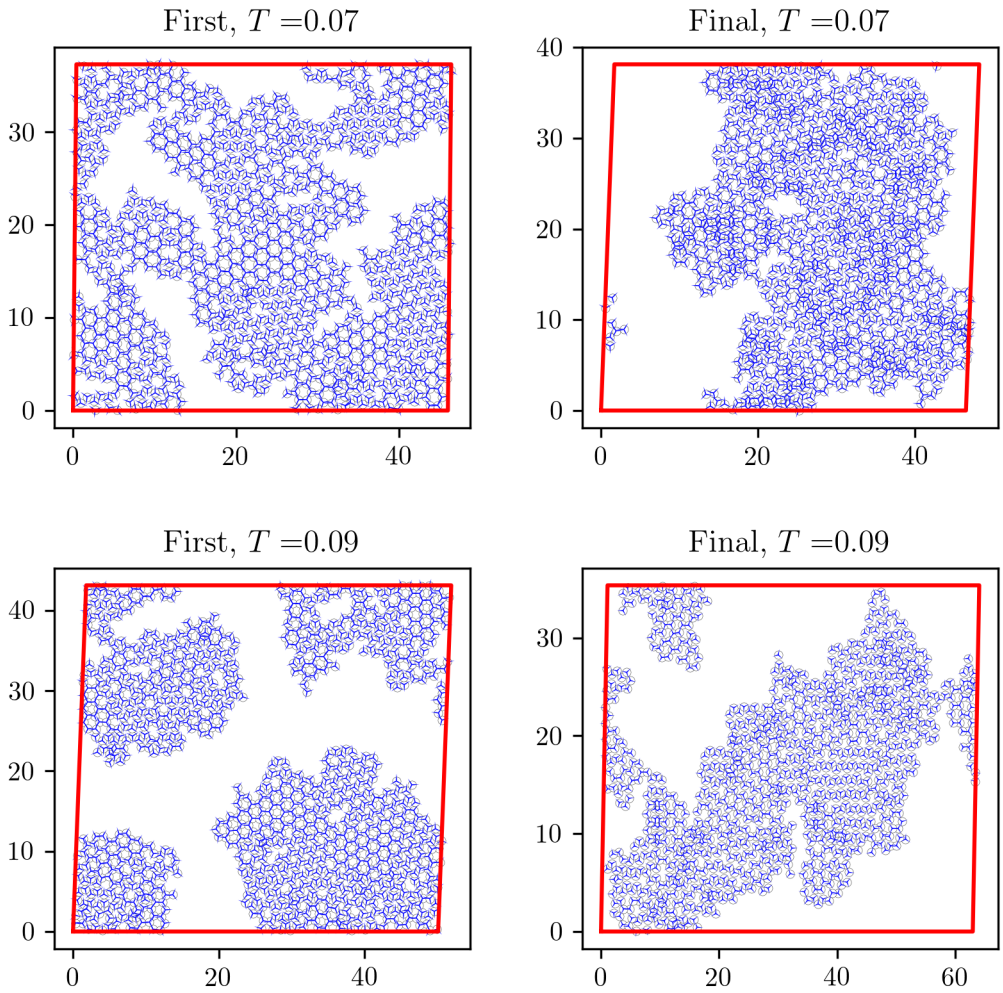

Figure 5.11: The first and final frames are shown for the trajectories in Figure 5.9, at each respective temperature as indicated by the title of each panel.

Chapter 6

Generalised Infinite Switch Simulations

6.1 Introduction

It should at this point be clear that sampling of complex multi-modal energy landscapes can be accelerated by incorporating information from non-physical states. These types of accelerated sampling methods have been used by the Molecular Dynamics (MD) community for years, but recently, methods such as simulated tempering [52] have received an increasing interest within the Machine Learning community [2], whose sampling problems often are similar to those found in the MD. The work presented in Chapter 4 showed how it is possible to operate ST in the infinite switch limit and how this improves the efficiency of the sampling scheme.

This work focused exclusively on varying of the temperature using the potential energy as the collective variable. On the other hand, the infinite switch framework is flexible and lends itself to a straightforward generalization where general collective variables can be incorporated. For example, combining an infinite switch based sampling approach with a collective variable which has been learned for some data distribution could produce a powerful method. However, in this case one needs to be cautious regarding the variable used, such that it is known to accelerate the sampling under the infinite switch framework. This, as justified in Section 4.1.5, comes down to quantifying whether the system goes through a first order phase transition in the direction of the collective variable for some value in the tempering domain.

As these improved sampling schemes gain interest from a wider audience and are applied to problems that are not necessarily easy to interpret in terms of physical intuition, the abstraction of accelerated schemes becomes a valuable field of study. Indeed, quantifying the conditions under which these general accelerated sampling methods will perform well, becomes an important aspect of accelerated sampling methods. Under these circumstances, even, identifying the correct collective variable to apply the infinite switch sampling scheme becomes troublesome. One might therefore question the validity and relevance of using collective and extensive variables with commonly known physical interpretations. It could be argued that these problems should find their own set of variables, that perhaps represent a more sensible choice on which to base an infinite switch accelerated sampling method.

In this final chapter we explore some of the limitations of using more general collective variables and numerically illustrate the importance of the knowledge of phase transitions in these collective variables in order to accelerate the sampling. We also

explore the idea of using the knowledge of two collective variables and suggest a situation in which such methods could become useful. Finally, we couple the infinite switch methods with the barostat method presented in the previous chapter – which produces a set of methods that can be used to temper pressure, temperature or pressure and temperature simultaneously. As all these methods are general applications of the scheme derived in Chapter 4, we only state a minimal amount of detail and refer to that chapter and the paper by Martinsson et al. [31] for further explanation.

6.2 Generalised Infinite Switch Simulation

Consider a system with variables $(\boldsymbol{q}, \boldsymbol{p}) \in \Omega \times \mathbb{R}^n$ on the torus, $\Omega = \mathbb{T}^n$, with potential $U : \Omega \to \mathbb{R}$, smooth. Given some importance sampling function $\theta : \Omega \to \mathbb{R}$ and magnitude λ, we define the extended system $(\boldsymbol{q}, \boldsymbol{p}, \lambda)$, with Hamiltonian $H(\boldsymbol{q}, \boldsymbol{p}) = \frac{1}{2}\boldsymbol{p}^T M^{-1}\boldsymbol{p} + U(\boldsymbol{q})$ and mass matrix M, with invariant measure with density,

$$\rho(\boldsymbol{q}, \boldsymbol{p}, \lambda) = Z^{-1} w(\lambda) \exp\left[-\beta H(\boldsymbol{q}, \boldsymbol{p}) + \lambda\theta(\boldsymbol{q})\right], \tag{6.1}$$

for $Z < +\infty$ given $w(\lambda) > 0$ and $w(\lambda) \in L^1(\mathbb{R})$. The normalization factor is given by,

$$Z = \sqrt{\det M}\left(2\pi\beta^{-1}\right)^{D/2} \int_{\mathbb{R}} \int_{\Omega} w(\lambda) e^{-\beta U(\boldsymbol{q}) + \lambda\theta(\boldsymbol{q})} \,\mathrm{d}\boldsymbol{q}\,\mathrm{d}\lambda. \tag{6.2}$$

Langevin dynamics is employed to sample $(\boldsymbol{q}, \boldsymbol{p})$ and overdamped Langevin for λ,

$$\begin{aligned}
\mathrm{d}\boldsymbol{q} &= M^{-1}\boldsymbol{p}\,\mathrm{d}t, \\
\mathrm{d}\boldsymbol{p} &= -\nabla U(\boldsymbol{q})\,\mathrm{d}t + \beta^{-1}\lambda\nabla\theta(\boldsymbol{q})\,\mathrm{d}t - \gamma M^{-1}\boldsymbol{p}\,\mathrm{d}t + \sqrt{2\beta^{-1}}\,\mathrm{d}W, \\
\mathrm{d}\lambda &= \varepsilon^{-1}\beta^{-1}\theta(\boldsymbol{q})\,\mathrm{d}t - \varepsilon^{-1}\beta^{-1}\partial_\lambda \log w(\lambda)\,\mathrm{d}t + \sqrt{2\beta^{-1}\varepsilon^{-1}}\,\mathrm{d}\bar{W}.
\end{aligned} \tag{6.3}$$

Here, ε is the time-separation scaling parameter introduced such that the system can be averaged using standard homogenization techniques [83]. This argument was also outlined in terms of simulated tempering in Chapter 4. Note that this dynamics corresponds exactly to ISST with $\theta(\boldsymbol{q}) = U(\boldsymbol{q})$ and $\lambda = \beta - \beta_t$ where β_t is the continuous *tempering* variable. The infinite switch limit, $\varepsilon \to 0$, of this dynamics has optimal convergence [31] and is implemented as,

$$\begin{aligned}
\mathrm{d}\boldsymbol{q} &= M^{-1}\boldsymbol{p}\,\mathrm{d}t, \\
\mathrm{d}\boldsymbol{p} &= -\nabla U_{\mathrm{e}}(\boldsymbol{q})\,\mathrm{d}t - \gamma M^{-1}\boldsymbol{p}\,\mathrm{d}t + \sqrt{2\beta^{-1}}\,\mathrm{d}W,
\end{aligned} \tag{6.4}$$

with the effective potential,

$$U_{\mathrm{e}}(\boldsymbol{q}) = U(\boldsymbol{q}) - \beta^{-1}\log\int_{\mathbb{R}} w(\lambda) e^{\lambda\theta(\boldsymbol{q})}\,\mathrm{d}\lambda. \tag{6.5}$$

The ergodic measure of (6.4) is the marginal distribution in $(\boldsymbol{q}, \boldsymbol{p})$ of (6.1), which has a density given by,

$$\bar{\rho}(\boldsymbol{q}, \boldsymbol{p}) = Z^{-1}\exp\left[-\beta\left(\tfrac{1}{2}\boldsymbol{p}^T M^{-1}\boldsymbol{p} + U_e(\boldsymbol{q})\right)\right]. \tag{6.6}$$

This approach is formulaic and these equations can be used to derive an array of methods utilizing different types of collective variables $\theta(\boldsymbol{q})$ and weight functions, $w(\lambda)$. In general, the choice of collective variable $\theta(\boldsymbol{q})$ should be informed by the desired

properties of the resulting method, but can also be learned [118, 119, 120].

We now make an important remark regarding the known issue of the lack of a continuous first order derivative of the free energy of the collective variable, which was discussed in detail in Section 4.1.5 and implies that the energy distribution cannot be made uniform in this range. As a consequence of this, it is not possible to infer that the infinite switch sampling will be accelerated.

Let $\theta(\boldsymbol{q})$ be a known collective variable. Define the free energy of this collective variable as,

$$F(z,\lambda) = -\beta^{-1}\log\int_{\Omega} e^{\lambda\theta(\boldsymbol{q})}\delta(z-\theta(\boldsymbol{q}))\,\mathrm{d}\boldsymbol{q}. \tag{6.7}$$

To make the statement above more precise, we note that the sampling will be accelerated by $\theta(\boldsymbol{q})$ if $\partial_\lambda F(z,\lambda)$ is continuous $\forall\lambda\in[\lambda_{\min},\lambda_{\max}]$. This is also conditional on the weight function $w(\lambda)$ being proportional to $\exp\left[-S(\lambda)\right]$ where,

$$S\left(\lambda\right) = \log\int_{\Omega} e^{-\beta U(q)+\lambda\theta(\boldsymbol{q})}\,\mathrm{d}\boldsymbol{q}. \tag{6.8}$$

We also draw the reader's attention to the fact that this has a physical interpretation and we say that there are no first order phase-transitions for any value of $\lambda\in[\lambda_{\min},\lambda_{\max}]$. This condition is trivial to check in simple cases where the closed form of the collective variable is known, but less intuitive for collective variables that are learned. Below, we work only in the domain of known collective variables and we invite future research into the tempering of learned general collective variables.

Let us state an equivalent form of equation (4.29):

$$\mathbb{E}_\lambda\left[\varphi\right] = Ze^{-S(\lambda)}\int_{\mathbb{R}}\int_{\mathcal{D}}\varphi(\boldsymbol{q},\boldsymbol{p})\frac{e^{\lambda\theta(\boldsymbol{q})}}{\int_{\mathbb{R}} w(\lambda')e^{\lambda'\theta(\boldsymbol{q})}\,\mathrm{d}\lambda'}\bar{\rho}(\boldsymbol{q},\boldsymbol{p})\,\mathrm{d}\boldsymbol{q}\,\mathrm{d}\boldsymbol{p}. \tag{6.9}$$

This equation plays the same role as (4.29) does in the simulated tempering case which implies that it can also be used to derive the general case learning of $\exp[S(\lambda)]$.

Double Infinite Switch Simulation

An interesting use case for infinite switch simulations exists in the reinterpretation of the scheme as a form of parameter exploration scheme. In comparison to methods such as adaptive biasing force (ABF) (see e.g [66]) that only accelerate sampling between two metastable states, an infinite switch simulation generates a significant amount of data per trajectory – which is accessible via the importance weights. In contrast to traditional techniques such as replica exchange and simulated tempering, an infinite switch simulation does not suffer from a significantly increased computational cost as the number of interpolation points is increased i.e. the number of levels of λ. This implies that in the rare cases where there is no need to accelerate the sampling, an infinite switch simulation can be used to explore all values of the tempering parameter λ.

On the other hand, assume that we posess the knowledge of a "good" collective variable, $\theta_1(\boldsymbol{q})$, for which we know that $\partial_{\lambda_1}F_1(z,\lambda_1)$ is continuous for all values $\lambda_1\in\left[\lambda_{\min,1},\lambda_{\max,1}\right]$. The collective variable $\theta_1(\boldsymbol{q})$ will therefore accelerate the sampling. Introduce now a second collective variable $\theta_2(\boldsymbol{q})$ which has a free energy with a non-continuous first order derivative for some value of λ_2 in $\left[\lambda_{\min,2},\lambda_{\max,2}\right]$. The question which we pose in this scenario is: Can $\theta_1(\boldsymbol{q})$ be used to accelerate the sampling of $\theta_2(\boldsymbol{q})$?

In this case the derivation of the equations are again formulaic and we write,

$$\begin{aligned} \mathrm{d}\boldsymbol{q} &= M^{-1}p\,\mathrm{d}t, \\ \mathrm{d}\boldsymbol{p} &= -\nabla U_{\mathrm{e}}(\boldsymbol{q})\,\mathrm{d}t - \gamma M^{-1}\boldsymbol{p}\,\mathrm{d}t + \sqrt{2\beta^{-1}}\,\mathrm{d}W, \end{aligned} \tag{6.10}$$

where,

$$U_{\mathrm{e}}(\boldsymbol{q}) = U(\boldsymbol{q}) - \beta^{-1}\log\int_{\lambda_{\min,1}}^{\lambda_{\max,1}}\int_{\lambda_{\min,2}}^{\lambda_{\max,2}} w(\lambda_1,\lambda_2)e^{\lambda_1\theta_1(\boldsymbol{q})+\lambda_2\theta_2(\boldsymbol{q})}\,\mathrm{d}\lambda_1\lambda_2. \tag{6.11}$$

Note that the weight $w(\lambda_1,\lambda_2)$ is in general a non-separable function, which we could assume to be separable. However, this is not beneficial if we want to apply the adaptive learning strategy from Section 4.1.6. If this is the case, we assume that $w(\lambda_1,\lambda_2)$ is a non-separable function and we write the equivalent to (4.29) as,

$$\mathbb{E}_{\lambda_1,\lambda_2}\left[\varphi\right] = Ze^{-S(\lambda_1,\lambda_2)} = \int_{\mathbb{R}}\int_{\mathcal{D}} \frac{\varphi(\boldsymbol{q},\boldsymbol{p})e^{\lambda_1\theta_1(\boldsymbol{q})+\lambda_2\theta_2(\boldsymbol{q})}}{\int\int w(\lambda_1',\lambda_2')e^{\lambda_1'\theta_1(\boldsymbol{q})+\lambda_2'\theta_2(\boldsymbol{q})}\,\mathrm{d}\lambda_1'\,\mathrm{d}\lambda_2'}\bar{\rho}(\boldsymbol{q},\boldsymbol{p})\,\mathrm{d}\boldsymbol{q}\,\mathrm{d}\boldsymbol{p}. \tag{6.12}$$

Here, $\bar{\rho}(\boldsymbol{q},\boldsymbol{p})$ is the invariant measure associated with (6.10) i.e.,

$$\bar{\rho}(\boldsymbol{q},\boldsymbol{p}) = Z^{-1}\exp\left[-\beta\left(\tfrac{1}{2}\boldsymbol{p}^TM^{-1}\boldsymbol{p} + U_e(\boldsymbol{q})\right)\right], \tag{6.13}$$

and Z is a normalisation constant. These are double integrals and must be solved numerically in $d = 2$. This generates an increased computational cost, however, this cost is insignificant when compared to the force evaluation. Below, we illustrate the use of these equations numerically.

Barostat Based Infinite Switch Simulation

For a vast range of computational biology and chemistry, barostats are essential tools that underpin most numerical experiments. In this case, the effective position used is the pair $(\boldsymbol{q}, L)$ where we denote the volume of a square simulation cell as, $V = L^3$. These equations again result in force rescaled SDEs and it is straightforward to couple the resulting equations with the algorithm given in the Chapter 5. For clarity, we assume that the simulation cell is cubic in which case the infinite switch SDEs are written as,

$$\begin{aligned} \mathrm{d}\tilde{\boldsymbol{q}} &= M^{-1}L^{-2}\tilde{\boldsymbol{p}}, \\ \mathrm{d}L &= \mu^{-1}L_{\boldsymbol{p}}, \\ \mathrm{d}\tilde{\boldsymbol{p}} &= -\tilde{\nabla}U_e(\tilde{\boldsymbol{q}},L)\,\mathrm{d}t - \gamma M^{-1}\boldsymbol{p}\,\mathrm{d}t + \sqrt{2\beta^{-1}}\,\mathrm{d}W, \\ \mathrm{d}L_{\boldsymbol{p}} &= \frac{1}{4}L^{-3}\tilde{\boldsymbol{p}}^TM^{-1}\tilde{\boldsymbol{p}}\,\mathrm{d}t - \partial_L U_e(\tilde{\boldsymbol{q}},L) - \nu\mu^{-1}L_{\boldsymbol{p}}\,\mathrm{d}t + \sqrt{2\beta^{-1}}\,\mathrm{d}W. \end{aligned} \tag{6.14}$$

Here, the effective infinite switch potential is given by,

$$U_e(\boldsymbol{q},L) = U(L\tilde{\boldsymbol{q}}) + \mathcal{P}L^3 - \beta^{-1}\log\int w(\lambda)\exp\left[\lambda\theta(\tilde{\boldsymbol{q}},L)\right]\mathrm{d}\lambda, \tag{6.15}$$

and $\tilde{\nabla}$ denotes the gradient w.r.t. to $\tilde{\boldsymbol{q}}$, the non-dimensional position in the simulation cell.

It is of course possible to chose the tempering variable as the temperature and

implement the original ISST scheme for a barostat. A perhaps more interesting choice is the utilization of,

$$\lambda = \mathcal{P} - \mathcal{P}_t, \quad \theta(\tilde{\boldsymbol{q}}, L) = \beta L^3, \tag{6.16}$$

which naturally must be dubbed the "PISS" algorithm. Equivalently, the double tempering scheme utilizing this setup, becomes "PISST" where,

$$\theta_1(\tilde{\boldsymbol{q}}, L) = \beta L^3, \ \lambda_1 = \mathcal{P} - \mathcal{P}_t \quad \text{and} \quad \theta_2(\tilde{\boldsymbol{q}}, L) = U(L\tilde{\boldsymbol{q}}), \ \lambda_2 = \beta - \beta_t. \tag{6.17}$$

Observe that the temperature tempering affects both the volume and the particle dynamics whereas pressure tempering only affects volume dynamics.

6.3 Applied Field Tempering

We continue the discussion numerically and study a few examples of the schemes described above for a set of small test problems. In the first case, we illustrate the importance of the knowledge of any first order phase transitions in the collective variable and illustrate how a poor choice of tempering variable will impede the sampling. In the following subsection we illustrate the use of a double tempering scheme for the Curie Weiss model from Section 4.3.2. We finish the discussion by applying two tempering schemes using the NPT ensemble for the Mercedes Benz model.

6.3.1 Double Well

Consider the double well potential for $x \in \mathbb{R}$, with an added field tempering parameter i.e. $\lambda = b_t \in [-b_{\max}, b_{\max}]$,

$$U(x, b_t) = 4\left(x^2 - 1\right)^2 + b_t\left(x + 1\right). \tag{6.18}$$

This additional parameter introduces a tilting of the original potential $U(x, 0)$ and is illustrated for $b_t = -1$ and $b_t = 1$ in Figure 6.1. A simple calculation shows that the free energy of $(x + 1)$ in this case is not continuous in b_t since,

$$\frac{\partial U(x, b_t)}{\partial b_t} = \frac{b_t}{|b_t|}(x + 1). \tag{6.19}$$

This is discontinuous precisely at $b_t = 0$, and we stress that the potential (6.18) satisfies conditions under which an infinite switch in b_t is expected to perform poorly. A tempering scheme based on the collective variable $\theta(x) = (x + 1)$ results in an effective potential of the form (6.5), which in this case takes the explicit form,

$$U_e(x) = U(x, 0) - \beta^{-1} \log \int_{-b_{\max}}^{b_{\max}} \omega(b_t) e^{-b_t(x+1)} \, \mathrm{d}b_t. \tag{6.20}$$

Here U is given by (6.18) and as described in Section 4.1.5, we chose the weight in (6.20) as $\omega(b_t) \propto Z^{-1}(b_t)$ which is learned on-the-fly.

In the right panel of Figure 6.1 we illustrate the original potential in black and the effective potential (6.20) for $b_{\max} = 50$ in dashed blue. It is clear that the effective potential which we would sample in this case has a larger energetic barrier than the original potential and means that the method slows the sampling exploration instead of accelerating it.

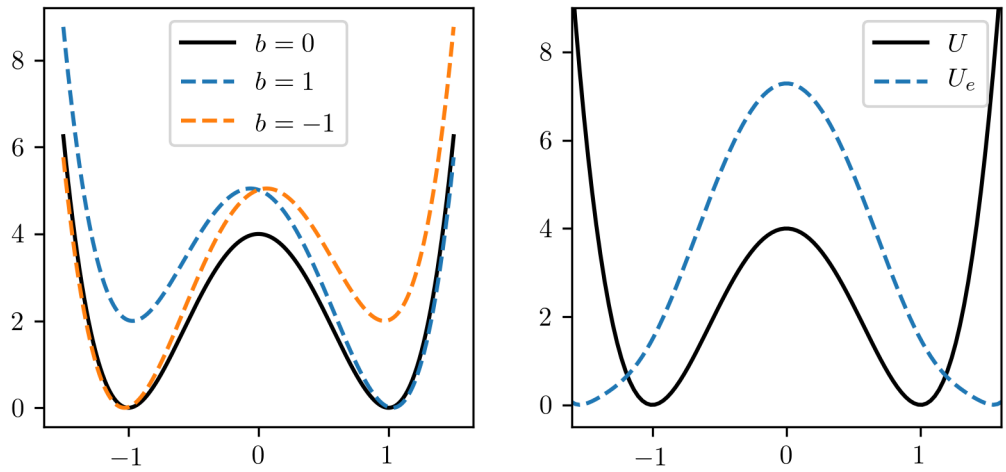

Figure 6.1: (Left) A plot of the tilting imposed by adding a linear term in the potential, given by (6.18). (Right) This plot shows the analytical form of (6.20) for $b_{\max} = 50$ and clearly illustrates the increase in the potential barrier emitted by this infinite switch scheme.

In Figure 6.2 we plot a sampled effective potential for a modest value of $b_{\max} = 2$ in orange, which agrees with the effective potential calculated using quadrature – in blue. This clearly indicates that the continuity condition on the first order derivative of the free energy of the collective variable, matters as the applied collective variable can completely reverse the desired effect and instead increase energetic barriers rather than decrease.

6.3.2 Curie Weiss Magnet

Consider again the Curie Weiss magnet with K spins and external magnetic field b,

$$U^K(m; b) = -K\left(\tfrac{1}{2}m^2 + bm\right). \tag{6.21}$$

Here, the magnetization m is defined as,

$$m = \frac{1}{K}\sum_{i=1}^{K} \cos\theta_i, \tag{6.22}$$

where $\theta_i \in [-\pi, \pi]$ is the angle of the i^{th} spin particle. We introduce the collective and tempering variables as the two pairs,

$$\theta_1(m; b) = U^K(m; b),\ \lambda_1 = \beta - \beta_t \quad \text{and} \quad \theta_2(m) = Km,\ \lambda_2 = (b - b_t). \tag{6.23}$$

The second tempering pair controls the strength and direction of the external field through b_t and the first pair is identical to ISST. It is known that ISST accelerates the exploration in magnetization for this model [31]. Identically to the previously discussed implementations, we adjust the weights $w(\lambda_1, \lambda_2)$ such that they are proportional to the correct partition function which is learned on-the-fly.

Under the assumption that we are using a sufficient number of spin particles, 15 or more, and a low enough temperature $T = 1/3$ this model cannot be sampled successfully

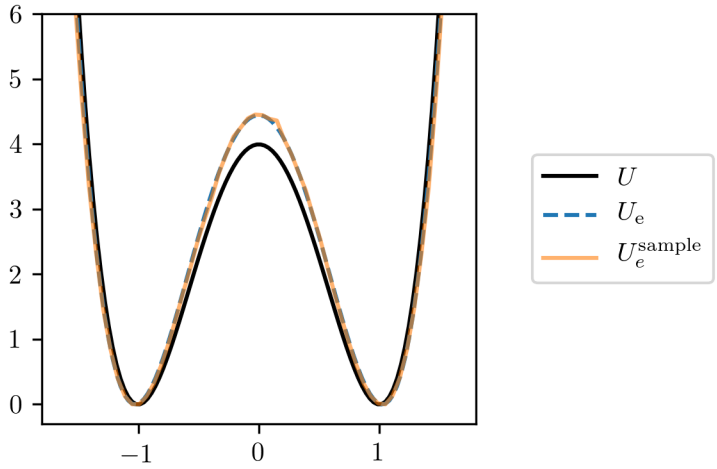

Figure 6.2: A plot of the sampling performed with an infinite switch scheme with a modest tempering domain $b_{\max} = 2$ and 15 interpolation points. The quadrature result is plotted in dashed blue with the orange plot representing the sampling results from an infinite switch trajectory.

using standard sampling techniques such as Langevin. From (6.21) it is clear that when $b = 0$ the model emits an even double well in magnetisation, which is tilted by tuning the externally applied magnetic field, b. The introduction of this field has the same characteristics as the tilting effect seen in the double well example from above, which first order free energy derivative in b_t is discontinuous at $b_t = 0$.

Let us introduce the premise of the numerical experiments. Assume that we want to collect statistics for $K = 25$ spins at the relatively cold temperature of $T = 1/3$, for which (6.21) is metastable in the magnetization m. Say that, additionally, we are interested in the response of the system as the external applied field, b_t, is changed from negative to positive. Tempering in this variable alone will not work for a standard simulated tempering scheme and we consequently propose to use a double tempering scheme using the pairs given in (6.23).

In this case we are free to choose the relative temperature β, which will affect the sampling performance significantly. A suitable choice of this parameter is the central point of the range $[\beta_{\min}, \beta_{\max}]$, as the level set explored at this value maximizes the support at both $\beta_{\min}$ and $\beta_{\max}$ (see Figure 4.4).

In Figure 6.3 we plot the effective free energy potential for three values, midpoint and the extremes in temperature. It is clear that the change in the height of the energetic barrier is significantly affected by the choice of reference temperature in this case.

In general it is difficult to predict what a suitable reference temperature should be for a single or double infinite switch tempering scheme when used to sample a biomolecule. However, a reference temperature chosen at the center of the tempering domain is most likely a suitable initial guess. We motivate this argument by pointing out that this is the level set which "has the most in common" with the neighbouring sets at either extrema, and should thus also support the least noisy collection of statistics over the entire range. In the experiment below we take $\beta = 1.5$.

When combined with trajectory reweighting, the double infinite switch simulation produces a large amount of data for every permutation of λ_1 and λ_2 in their respective

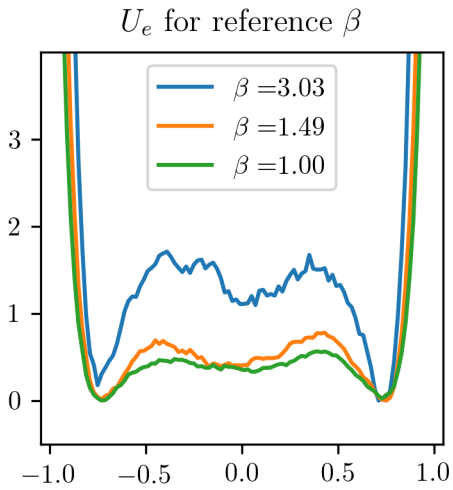

Figure 6.3: Shows the double tempering effective free energy, that is sampled for different choices of the reference temperature β given in the legend.

ranges. In the experiment displayed in Figure 6.4 we show the results of this exercise when applied to the Curie Weiss model with 25 spins. In this example we used 15 discrete values for both the applied field and temperature, which results in the production of 225 trajectories from a single double tempering trajectory, of which a subset is shown in Figure 6.4.

Additionally, this plot nicely illustrates the exact connection that is made between the two collective variables and utilises the bifurcation in magnetisation and temperature to connect the two metastable states in the low temperature regime. The energetic barrier can therefore be avoided by allowing the system to explore the temperature domain. We also remark that the double tempering works well in this case as the applied field effect vanishes in the high temperature limit. This is why the sampling in the lower temperature limit is excellent and the double tempering scheme performs very well for this model. We conclude that Figure 6.4 is encouraging and that the a double tempering scheme can perform well for certain problems.

6.3.3 Mercedes-Benz Potential

We first state that in this case it is not understood, at the time of writing, whether the system has a phase transition in the tempering variables exploited in this section. We therefore cannot guarantee that the application of an infinite switch algorithm in this case will improve the sampling performance of this model. With this in mind, we progress.

The Mercedes-Benz (MB) potential was described in Section 5.3.2 to which, and the references there, we refer for a more detailed description of this model. Our aim in this section is to discuss some of the possibilities and computational complexities, which we have not yet addressed and arise with the implementation of infinite switch simulation sampling schemes. In Section 4.3.4 we briefly touch upon issues regarding choice of collective variable e.g the entropic explosion of available configurations emerging in high temperature states. Below we extend this discussion in more detail using the Mercedes-Benz potential as an illustrative example.

The most decisive factor in the successful implementation of any infinite switch

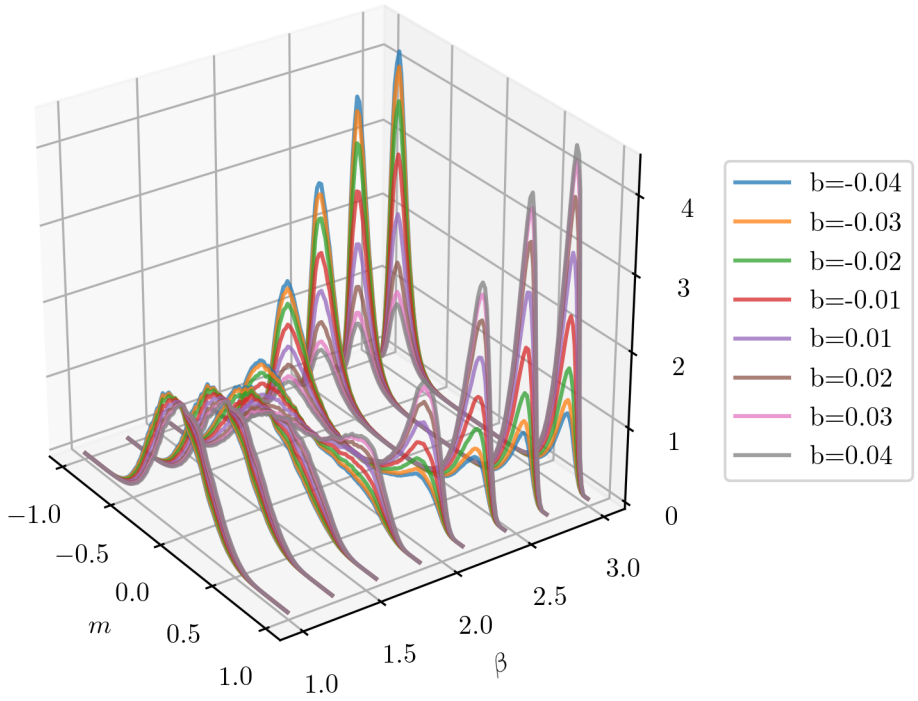

Figure 6.4: A $d = 3$ plot of the reweighted data from a double sampling trajectory of the Curie Weiss model with 25 spins and 15 interpolation points for each tempering variable, for a total of 225 (not all displayed) reweighted trajectories. Here $b = 0$ with $b_{\max} = 0.04$ and the reciprocal temperature $\beta = 2/3$ for $\beta_t \in [1, 3]$.

scheme stems from the need to calculate exponentials of the potential energy. This becomes troublesome as the systems grows in size and the absolute value of the potential energy increases in magnitude. The absolute value of this energy is irrelevant for many experiments and a favourable approach to counteract the evaluation of exponentials with large arguments is to introduce an arbitrary additive constant that minimises the argument. This works very well for small simple systems, where the potential energy is relatively well behaved. To further stabilise the evaluation of exponentials, especially for larger systems, we use the identity,

$$\log \sum_{i=0} a_i = \log a_0 + \log \left(1 + \sum_{i=1} \frac{a_i}{a_0} \right), \tag{6.24}$$

where necessary, ensuring to sort the coefficients a_i in descending order.

In the case of larger systems, such as the 1024 particle MB model considered in this section, the potential energy changes rapidly over the tempering domain and it is difficult to determine a suitable constant. This limits the usable range and the infinite switch simulations are constrained to a narrow band as otherwise the observable weights (6.9) become problematic to calculate as they are vanishingly small. This also destabilises the algorithm and we are forced to limit the exploration to a narrow window.

The experiments we conduct in this section derive from the experiments done in Section 5.3.2 and use the same timestep and parameters, with NPT sampling performed by the algorithm described in Chapter 5. We extend this scheme by first coupling it

with the PISS algorithm described at the end of Section 6.2 and secondly with the ISST algorithm. We use the final states displayed in Figure 5.8 (denoted as "Pressure 0.04" and "Pressure 0.08") as the initial conditions for the PISS and ISST experiment respectively. By matching either the target reference pressure or target reference temperature in these experiments we ensure that the experiments remain near equilibrium.

With the additional tweaking outlined just above, we use the generalised version of the adaptive learning process described in Section 4.1.6 and briefly outlined in Section 6.2, to adjust the tempering parameter weights $w(\lambda)$ on the fly. For each trajectory we also record the observable weights such that (6.9) can be used to reweight each infinite switch trajectory to find an observable for any given value in the tempering domain.

For each of the two infinite switch experiments we choose the respective target reference pressure and target temperature as the central point of the chosen tempering range. This should ensure that we maximise the overlap between all conditions. Our intention is also to push the limits of the algorithm and we therefore choose the range in each case near the limit for which the algorithm is stable. Together with the large number of particles we are approaching the numerical threshold for which the evaluation of exponentials are stable in `double` precision on a 64-bit machine. Because of this we emphasise that the results below should be considered with some caution, and if the aim is instead to obtain more reliable results the tempering range should be halved as a first step.

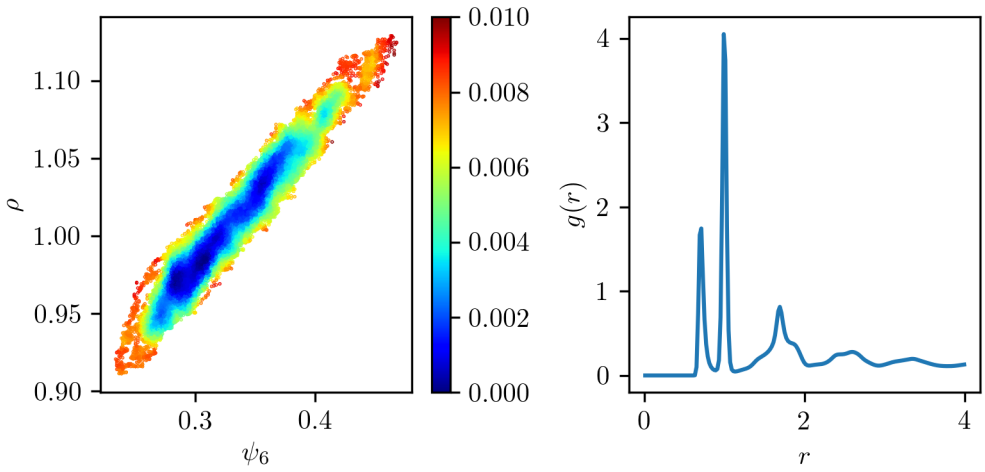

Figure 6.5: A heatmap illustrating the averaged free energy surface explored by the PISS algorithm in the left panel and its radial distribution to the right. Here the target temperature was $T = 0.05$ and the target pressure was set at $\mathcal{P} = 0.05$ with 25 tempering values in $[0.04, 0.06]$.

In a first experiment we use the PISS algorithm with 25 tempering pressures in $[0.04, 0.06]$ and a reference target pressure of $\mathcal{P} = 0.05$ at constant temperature $T = 0.05$. In the left panel of Figure 6.5 we plot the free energy of the effective potential U_e and the order parameter introduced in Section 5.3.2. As before we use ρ to again denote the number density. In the left panel of the same figure we show the radial distribution function, which drawing from Section 5.3.2 indicates that we are in a mixed state consisting both of the oblique and hexagonal crystal structures as indicated by the two tallest peaks. Similarly to what we saw in Section 5.3.2 we observe a wide response in the density as a result of the constant pressure, which is most likely caused by the

random decay and formation of denser subs structure such as the tetratic imperfections [114].

Another notable feature about the effective free energy panel is that it appears to sample conditions near $\mathcal{P} = 0.05,\ T = 0.05$ closely, not exploring other conditions much. This suggests that we cannot expect the reweighting to be successful as we essentially miss support for other conditions, by only exploring states that are relevant to $\mathcal{P} = 0.05,\ T = 0.05$. This can be concluded by using Figure 5.7 and realising that the densities explored in Figure 6.5 are precisely halfway in-between the conditions with target pressures set to $\mathcal{P} = 0.04$ and $\mathcal{P} = 0.06$.

This suggests that an intimate physical understanding or some prior exploration of the relevant, to the specific model, states are needed. In particular, it is crucial to understand the response of the system as the tempering variables are changed such that the chosen range produces sensible conditions with sufficient overlap between states at either extreme of the range. This is not the case in these experiments as we are exploring the limits of the algorithm.

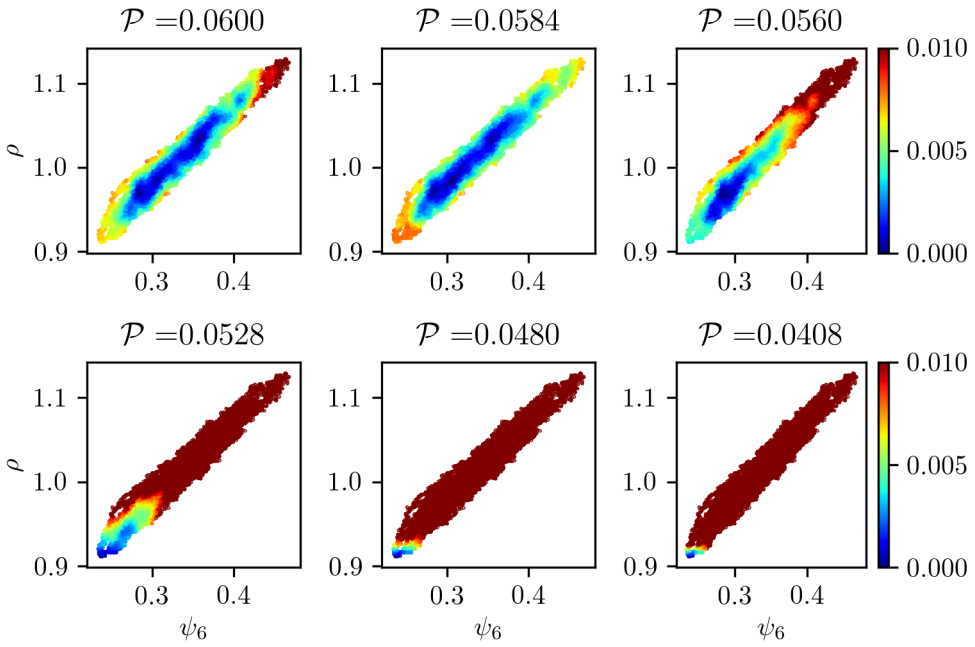

Figure 6.6: Reweighted heat maps are shown for 6 values of the pressure, showing the difficulties the algorithm has in exploring the phase space. The algorithm seem to capture the expected behaviour in the observable weights as lower density states are favoured as the pressure is decreased, for values that are in agreement with Figure 5.7.

The problem arising when using large tempering domain become apparent if we consider the reweighted free energies which are shown for six values in Figure 6.6. Comparing again to Figure 5.7 we observe that for $\mathcal{P} = 0.06$, the two figures do not agree. Observe that the results in Figure 5.7 were started in a $\rho = 0.7$ state and progresses into a higher density state ($\rho = 1.2$) during simulation. On the other hand, in Figure 6.6 the top left panel suggests that stable conditions at this pressure is for $\rho = 1.0$. We attribute this disagreement to the extreme conditions used for these latter experiments and it should serve as a reminder to proceed with caution as all panels in Figure 6.6 look believable, had we not possessed additional information.

Interestingly, in the lower pressure limit we seem to get agreement between Figure 5.7 and Figure 6.6 as the weights for these pressure values indicate that none or few of the sampled states are available at the, in the title, indicated target conditions. This is rather surprising and we speculate that it results from the more stable numerical evaluation of exponentials in this limit.

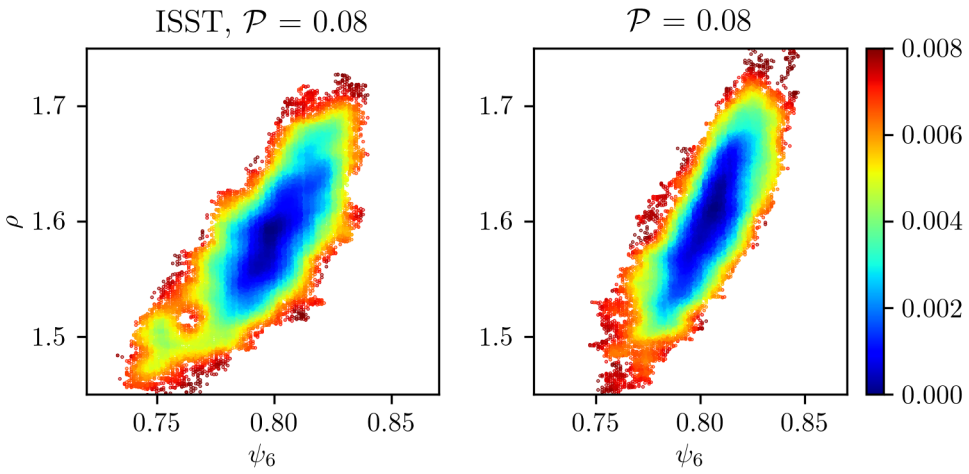

Figure 6.7: Plots are shown displaying the differences in the free energy explored when ISST is applied to the Mercedes-Benz potential. *(left)* A panel showing the effective free energy explored by ISST at constant target pressure 0.08 using the barostat from Chapter 5. In this simulation the reference target temperature was set to $\beta = 1/0.05$ and the tempering used 25 Legendre roots in the domain $[0.06, 0.04]$. *(right)* A panel containing a heat map of a trajectory starting from the same initial condition but sampled without using the ISST algorithm.

We continue the numerical investigation of the computational limit of the infinite switch schemes, using now instead the ISST algorithm. We first remark that this method is swapping free in temperature only using a single reference temperature to perform the experiments. This helps to stabilise the barostat algorithm which is only exposed to a single temperature, minimising the ringing which could occur as a result of using standard simulated tempering.

In Figure 6.7 we have plotted the results from standard Langevin sampling under the same target conditions with the same initial condition in the right panel. In the left panel of the same figure we plot the free energy of the effective potential U_e, sampled at the reference pressure and temperature of $\mathcal{P} = 0.08,\ T = 0.05$. It is clear that the free energies in this case have a better overlap in the density than in the previous case as a result of the pressure being kept constant.

In the panels in Figure 6.8 we plot six reweighted free energies using ISST. As in the previous case we observe that the algorithm struggles to determine the weights in one limit and we note that it is unlikely that the top row in Figure 6.8 is correct. However, the lower row is interesting and building on the result of the previous experiment we assume that these values are correct. Under this assumption we see an emergence of a large number of metastable regions as the temperature is decreased, with the largest well concentrated at higher densities as expected. It seems to suggest that the model is very sensitive to both the initial condition and target temperature and pressure imposed.

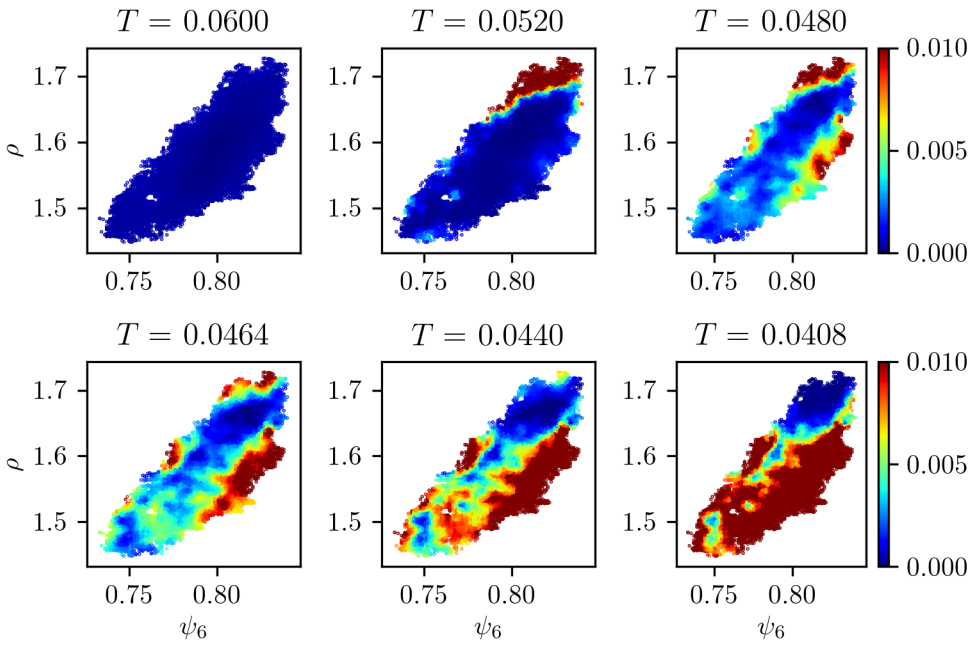

Figure 6.8: Reweighted heat maps are shown for six values of the temperature and display the difficulties the algorithm has in determining the observable weights for larger systems. Note that these heat maps should not be trusted to be correct and are only displayed in order to outline difficulties in applying the algorithm. However, at the lower temperatures on the bottom row we see an indication of the expected concentration in higher density states and order as the temperature is decreased.

In particular the final frame with a target temperature of $T = 0.0408$ seem to emit a low density metastable state at $\rho = 1.5$ and $\psi_6 = 0.75$. This state would be difficult to find using standard simulation methods as it would require starting with a very specific initial condition near these particular values as the state appears largely isolated. Increasing the temperature from this state, focusing on the panel labeled $T = 0.0440$, we observe a rapid response to the change in temperature as the metastable states appear to connect. The previously isolated state increases in size and appears to exhibit a bi-modal structure.

This final experiment illustrates the problems arising with the use of annealing techniques (Section 3.1), as often done for these models, since it is entirely possible that these algorithms get stuck in some low energy states largely isolated from other perhaps more viable states. It also makes a strong case for the need of barostats with precise pressure and temperature control as these isolated states could likely be destabilised by methods with poor pressure control, resulting in them being overlooked by practitioners. Finally, we do note that the conditions sampled with the algorithm only explores states that are mostly relevant to the reference conditions, here $\mathcal{P} = 0.08$, $T = 0.05$, also limiting the exploration of phase-space relevant to other conditions in this case.

We conclude this section by remarking on the difference in the two final states obtained by the standard Langevin and ISST trajectories, plotted in Figure 6.9. Comparing the two frames we see a greater alignment in the crystal structure as a result of using the ISST algorithm. In contrast to this, in the standard sampling scheme's final frame we observe configurations aligned in the hexagonal phase, instead of the oblique

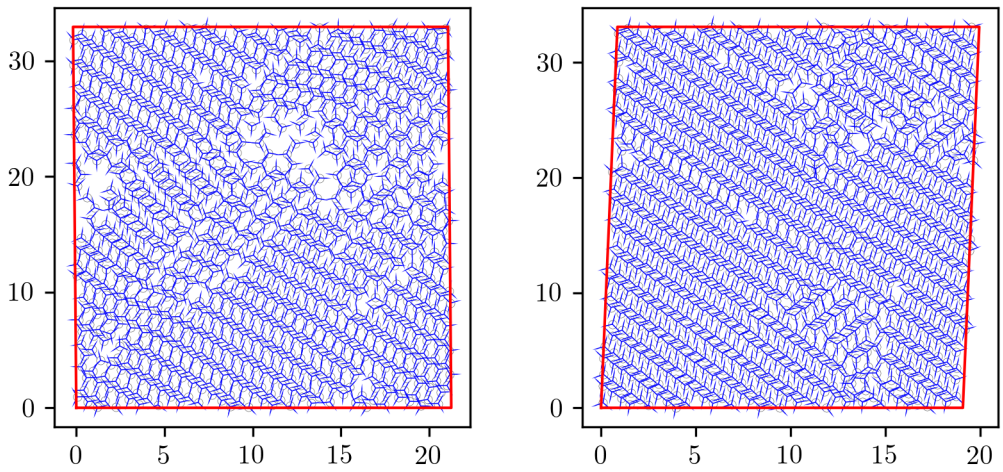

Figure 6.9: Two panels showing (left) the final frame obtained when sampling without ISST and then the final (right) state when sampling with ISST. Note the removal of the elongated defect in the right panel of the final configuration.

state as we expect (and also observe a majority of) under higher pressure conditions. This suggests that there is a structural difference between the two experiments and when using ISST we have, in this case, favoured a condition which is expected at lower temperatures.

Chapter 7

Conclusion

In this thesis we have presented work on accelerated sampling methods that improve the sampling via phase-space extension. In Chapter 1 we introduced concepts from statistical mechanics, such as thermodynamics, and we made the connection with sampling of more general statistical models such as Bayesian inference. We also introduced sampling in a broader context with the application in quantifying the fairness of geopolitical redistricting i.e. gerrymandering. This set the scene for the broad area of different applications where sampling finds use and illustrated some of the challenges needed to be addressed.

The introductory chapter was followed by a short overview of terms from stochastic analysis in Chapter 2 that formalises many concepts and measures that are used to quantify sampling performance of stochastic numerical methods. In the final two sections of this chapter we introduced two common approaches to sampling, which we concluded with an overview on how schemes with desirable properties are constructed for SDEs.

In Chapter 3 we outlined the chronological development of accelerated methods relevant to this thesis. We started our overview describing simulated annealing as a way to optimise complex problems remarking on that it is not a sampling method. This led to the development of simulated tempering which is a method that extends phase space by constructing a random walk over some pre-defined temperature range. Finally, we discussed the replica exchange method and the theoretical studies undertaken in this context that underlie our own work, presented in the following chapter.

The first chapter to contain novel work undertaken during the course of the PhD is in Chapter 4. We rewrote the presentation of the introduction and foundations of [31] to both highlight and simplify the presentation of some of the important concepts from the original paper. We showed that it is optimal to operate simulated tempering in the infinite switch limit for which we write down a dynamical equation, that is both simple to implement and to couple with an adaptive recurrence relation for adjusting the unknown temperature weights. We also outlined some of the limitations and made predictions for situations under which the method's performance is not guaranteed, which we showed can be quantified in terms of a first order phase transition. We ended this chapter with the numerical experiments from the original paper.

With a change of topic in Chapter 5, we presented work on the derivation of a robust and accurate constant pressure Langevin based thermostat. This chapter does not contain any material that relates to accelerated sampling but came about as a natural need for a Langevin based barostat, on which we could build accelerated sampling schemes for NPT. We followed previous work done in the context of NVT sampling

and carefully considered areas in which existing literature was lacking detail. From this we derived a set of schemes that were investigated numerically and the excellent properties of these schemes are illustrated in the final section of this chapter.

In the thesis concluding Chapter 6 we numerically investigate the properties and limitations of generalized infinite switch schemes. We explicitly show that these methods do not perform well for systems where first order phase transitions are present, although if we possess knowledge of a second collective variable with no known phase transition, the effects can be remedied. We finally discussed some of the issues associated with the implementation of these types of schemes and explicitly outline some of the considerations we made in order to achieve good performance of the algorithms. These types of implementation issues are particularly noticeable in large scale systems and we illustrated some areas of concern that need to be addressed going forward with new research in this area.

In particular, it is clear in the final section of Chapter 6 that the main complication in the implementation and execution of infinite switch algorithms is the need to evaluate terms of the form $\exp[\lambda\theta(x)]$. When evaluating such terms of extensive collective variables (collective variables that grow with system size), the available range of the tempering variable λ i.e. $[\lambda_{\min}, \lambda_{\max}]$ must be made sufficiently small in order for the resulting algorithm to be stable. This is problematic since the acceleration in the method, to a certain extent, derives from the use of a large range of values; this is of course very problem dependent and not always the case. As we mentioned in Section 6.3.3 these effects can be controlled by introducing some constant that minimises the argument of the exponential and functions well when energy fluctuations are small.

Interesting avenues exist for further research into ensuring the stability of generalised infinite switch algorithms when used with larger tempering domains $[\lambda_{\min}, \lambda_{\max}]$. A first step in this direction that could find use with crystals or single particle materials, would be to use the average $\theta(x)$ per particle instead of the full collective variable. This would easily control problems arising from the evaluation of exponentials but would likely limit the usefulness of the algorithm as the force is only modified by an average. Locally this could be problematic as substructures can form that require a more aggressive modification to unravel which means that the algorithm will not be as effective.

Tempering of extensive collective variables for MD has existed for a long time. The derivation of these schemes has been aided by physical intuition that guided the identification of suitable collective variable candidates used to base the schemes on. An example of this approach was presented in Chapter 6 where we introduce accelerated sampling schemes based on tempering in pressure with volume playing the role of the collective variable. The idea behind this scheme was to essentially allow for a small permutation in pressure near the target pressure that relaxes the imposed target pressure on the model, allowing the molecule "more room" in which to spread out. By then utilising the importance weights collected along the infinite switch trajectory, conditions at the physically relevant state can be inferred via reweighting. These types of general tempering schemes could find creative solutions to problems currently only confronted with temperature tempering and we hope that this initial exploration encourages further investigation into the design of generalised infinite switch simulation schemes.

As computational power increases and sampling finds use in an increasing number of applications reaching beyond traditional areas such as statistical physics, the need for abstract accelerated sampling schemes will grow. Understanding the limitations and conditions under which these schemes operate well become increasingly important and

must be addressed by abstract statements. As an example consider ISST. Heuristically it is straightforward to understand why and how this scheme works in MD (e.g for a protein) as an increase in temperature implies that more energy is added to the system, allowing the molecule to unravel from some complicated folded state. The same holds true for tempering of an applied field or pressure which derivation was also motivated by a physical understanding. For general infinite switch tempering methods applicable to statistical models we do not have the same physical understanding and as we saw in the case of gerrymandering, it is difficult to both make predictions of what collective variables will successfully temper the system, or indeed understand why temperature works poorly in this case.

This brings our discussion to collective variables and the role they have in generalised infinite switch sampling for general statistical models, not necessarily easily understood through classical statistical mechanics. Designing or otherwise finding collective variables that accelerate these sampling problems is an interesting topic for further research which we have aimed at initially stimulating in the final parts of the thesis. For these cases, it is clear that a more creative approach is needed and that the tempering of temperature does not posses the right tempering qualities or is simply too aggressive (explosion of available states at higher temperature). Indeed, as we saw in Section 6.3 it is possible to define two collective variables for the same systems: one which works well with no first order phase transitions and a second that exhibits a phase transition. This encourages the study of collective variables for systems where first order phase transitions in temperature are present. Assuming that such variables exist without first order phase transition in the tempering range, we introduce a new generalised infinite switch simulation schemes which can be used to explore the system. We also introduced the use of a double tempering which could also be used in this case.

Automatically detecting such variables for particular systems could lead to a large number of generalised infinite switch simulation methods that are applicable for a large number of models. This is a difficult problem and even defining the criteria such collective variables should satisfy, is non trivial. Finding the design criteria of collective variables that will result in infinite switch sampling schemes with good performance is an interesting topic for future research. More understanding is needed in this direction, both to understand how to define their derivation but perhaps more urgently, understand situations when the tempering of a particular collective variable will drive the exploration of some model well.

Collective variables are a generalisation of reaction coordinates, but in practice, it is not always clear how to use the added flexibility. As an example consider the adaptive biasing force (ABF) method [66] which eliminates free energy barriers along some reaction coordinate. This method performs exceedingly well when coupled with the field tempering variable b_t in Section 6.3 and completely eliminates the barrier between the metastable states. The difference and similarities between infinite switch simulated tempering and using β as the reaction coordinate in ABF are not known at this point.

In this thesis we have presented some temperature and pressure based accelerated sampling schemes and some evidence for the need of further research into the design of collective variables for generalised infinite switch simulation schemes. Theoretical advances into the understanding and quantification of conditions that such collective variables should satisfy is needed and would ease the adoption of such techniques beyond areas of statistical physics or MD. If such conditions can be quantified and coupled with adaptive learning, the impact on the wider sampling community would be significant.

Bibliography

[1] G. A. Ross, A. S. Rustenburg, P. B. Grinaway, J. Fass, and J. D. Chodera, "Biomolecular simulations under realistic macroscopic salt conditions," *The Journal of Physical Chemistry B*, vol. 122, no. 21, pp. 5466–5486, 2018.

[2] L. Leng, R. Martel, O. Breitwieser, I. Bytschok, W. Senn, J. Schemmel, K. Meier, and M. A. Petrovici, "Spiking neurons with short-term synaptic plasticity form superior generative networks," *Scientific Reports*, vol. 8, no. 1, pp. 1–11, 2018.

[3] G. Herschlag, H. S. Kang, J. Luo, C. V. Graves, S. Bangia, R. Ravier, and J. C. Mattingly, "Quantifying gerrymandering in North Carolina," *arXiv:1801.03783*, 2018.

[4] F. Fratev and S. Sirimulla, "An improved free energy perturbation fep + sampling protocol for flexible ligand - binding domains," *Scientific Reports*, vol. 9, no. 1, pp. 1–13, 2019.

[5] C. Zannoni, "Molecular design and computer simulations of novel mesophases," *Journal of Materials Chemistry*, vol. 11, no. 11, pp. 2637–2646, 2001.

[6] R. Bellman, *Dynamic Programming*. Dover Books on Computer Science Series, Dover Publications, 2003.

[7] T. Lelièvre, G. Stoltz, and M. Rousset, *Free Energy Computations: a Mathematical Perspective*. London ; Hackensack, N.J: Imperial College Press, 2010.

[8] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff, "UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations," *Journal of the American Chemical Society*, vol. 114, no. 25, pp. 10024–10035, 1992.

[9] C. Kramer, A. Spinn, and K. R. Liedl, "Charge anisotropy : Where atomic multipoles matter most," *Journal of Chemical Theory and Computation*, vol. 10, no. 10, pp. 4488–4496, 2014.

[10] S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, "Towards exact molecular dynamics simulations with machine-learned force fields," *Nature Communications*, vol. 9, no. 1, pp. 1–10, 2018.

[11] C. Zanette, C. C. Bannan, C. I. Bayly, J. Fass, M. K. Gilson, M. R. Shirts, J. D. Chodera, and D. L. Mobley, "Toward learned chemical perception of force field typing rules," *Journal of Chemical Theory and Computation*, vol. 15, no. 1, pp. 402–423, 2019.

[12] B. Leimkuhler and S. Reich, *Simulating Hamiltonian Dynamics*. Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, 2005.

[13] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration: Structure-preserving Algorithms for Ordinary Differential Equations*. No. 31 in Springer Series in Computational Mathematics, Berlin ; New York: Springer, 2nd ed ed., 2006.

[14] J. C. Mattingly, A. M. Stuart, and D. J. Higham, "Ergodicity for SDEs and approximations: locally lipschitz vector fields and degenerate noise," *Stochastic Processes and their Applications*, vol. 101, no. 2, pp. 185–232, 2002.

[15] K. C. Zygalakis, "On the existence and the applications of modified equations for stochastic differential equations," *SIAM Journal on Scientific Computing*, vol. 33, no. 1, pp. 102–130, 2011.

[16] B. Leimkuhler and C. Matthews, "Rational construction of stochastic numerical methods for molecular sampling," *Applied Mathematics Research eXpress*, p. abs010, 2012.

[17] A. Abdulle, G. Vilmart, and K. C. Zygalakis, "High order numerical approximation of the invariant measure of ergodic SDEs," *SIAM Journal on Numerical Analysis*, vol. 52, no. 4, pp. 1600–1622, 2014.

[18] G. A. Pavliotis, *Stochastic processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin equations*, vol. 60 of *Texts in Applied Mathematics*. New York: Springer, 2014.

[19] A. Abdulle, G. Vilmart, and K. C. Zygalakis, "Long time accuracy of Lie–Trotter splitting methods for langevin dynamics," *SIAM Journal on Numerical Analysis*, vol. 53, no. 1, pp. 1–16, 2015.

[20] T. Lelièvre and G. Stoltz, "Partial differential equations and stochastic methods in molecular dynamics," *Acta Numerica*, vol. 25, pp. 681–880, 2016.

[21] R. Balian, *From Microphysics to Macrophysics: Methods and Applications of Statistical Physics*, vol. 1 of *Texts and Monographs in Physics*. Berlin ; New York: Springer-Verlag, 1991.

[22] R. Balian, *From Microphysics to Macrophysics: Methods and Applications of Statistical Physics*, vol. 2 of *Texts and Monographs in Physics*. Berlin ; New York: Springer-Verlag, 1991.

[23] R. Zwanzig, "Nonlinear generalized langevin equations," *Journal of Statistical Physics*, vol. 9, no. 3, pp. 215–220, 1973.

[24] L. Rey-Bellet, "Ergodic properties of Markov processes," p. 39.

[25] B. Leimkuhler and M. Sachs, "Ergodic properties of quasi-Markovian generalized Langevin equations with configuration dependent noise and non-conservative force," *arXiv:1804.04029*, 2018.

[26] H. C. Andersen, "Molecular dynamics simulations at constant pressure and/or temperature," *The Journal of Chemical Physics*, vol. 72, no. 4, pp. 2384–2393, 1980.

[27] D. V. Lindley, "The philosophy of statistics," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 49, no. 3, pp. 293–337, 2000.

[28] L. Verlet, "Computer " experiments " on classical fluids . i . thermodynamical properties of Lennard-Jones molecules," *Physical Review*, vol. 159, no. 1, pp. 98–103, 1967.

[29] B. Leimkuhler, M. Sachs, and G. Stoltz, "Hypocoercivity properties of adaptive langevin dynamics," *arXiv:1908.09363*, 2019.

[30] G. Vilmart, "Postprocessed integrators for the high order integration of ergodic SDEs," *SIAM Journal on Scientific Computing*, vol. 37, no. 1, pp. A201–A220, 2015.

[31] A. Martinsson, J. Lu, B. Leimkuhler, and E. Vanden-Eijnden, "The simulated tempering method in the infinite switch limit with adaptive weight learning," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, no. 1, p. 013207, 2019.

[32] P. Dupuis, Y. Liu, N. Plattner, and J. Doll, "On the infinite swapping limit for parallel tempering," *Multiscale Modeling & Simulation*, vol. 10, no. 3, pp. 986–1022, 2012.

[33] I. Bethune, R. Banisch, E. Breitmoser, A. B. K. Collis, G. Gibb, G. Gobbo, C. Matthews, G. J. Ackland, and B. J. Leimkuhler, "Mist : A simple and efficient molecular dynamics abstraction library for integrator development," *Computer Physics Communications*, vol. 236, pp. 224–236, 2019.

[34] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[35] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *The Journal of Physical Chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977.

[36] W. E, D. Liu, and E. Vanden-Eijnden, "Nested stochastic simulation algorithms for chemical kinetic systems with multiple time scales," *Journal of Computational Physics*, vol. 221, no. 1, pp. 158–180, 2007.

[37] W. K. Hastings, "Monte carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[38] D. Talay, "Stochastic hamiltonian systems : Exponential convergence to the invariant measure , and discretization by the implicit euler scheme," *Markov Processes and Related Fields*, vol. 8, no. 2, pp. 163–198, 2002.

[39] P. Rossky, J. Doll, and H. Friedman, "Brownian dynamics as smart monte carlo simulation," *The Journal of Chemical Physics*, vol. 69, no. 10, pp. 4628–4633, 1978.

[40] G. O. Roberts, A. Gelman, and W. R. Gilks, "Weak convergence and optimal scaling of random walk metropolis algorithms," *The Annals of Applied Probability*, vol. 7, no. 1, pp. 110–120, 1997.

[41] S. Nosé, "A unified formulation of the constant temperature molecular dynamics methods," *The Journal of Chemical Physics*, vol. 81, no. 1, pp. 511–519, 1984.

[42] W. G. Hoover, "Canonical dynamics: Equilibrium phase-space distributions," *Physical Review A*, vol. 31, no. 3, pp. 1695–1697, 1985.

[43] F. Legoll, M. Luskin, and R. Moeckel, "Non- ergodicity of the nos hoover thermostatted harmonic oscillator," *Archive for Rational Mechanics and Analysis*, vol. 184, no. 3, pp. 449–463, 2007.

[44] G. De Fabritiis, M. Serrano, P. Español, and P. Coveney, "Efficient numerical integrators for stochastic models," *Physica A: Statistical Mechanics and its Applications*, vol. 361, no. 2, pp. 429–440, 2006.

[45] N. Grønbech-Jensen and O. Farago, "A simple and effective Verlet-type algorithm for simulating Langevin dynamics," *Molecular Physics*, vol. 111, no. 8, pp. 983–991, 2013.

[46] N. Bou-Rabee and H. Owhadi, "Long- run accuracy of variational integrators in the stochastic context," *SIAM Journal on Numerical Analysis*, vol. 48, no. 1, pp. 278–297, 2010.

[47] B. Leimkuhler, C. Matthews, and G. Stoltz, "The computation of averages from equilibrium and nonequilibrium langevin molecular dynamics," *IMA Journal of Numerical Analysis*, p. dru056, 2015.

[48] A. Debussche and E. Faou, "Weak backward error analysis for SDEs," *SIAM Journal on Numerical Analysis*, vol. 50, no. 3, pp. 1735–1752, 2012.

[49] A. Rößler, "Stochastic taylor expansions for the expectation of functionals of diffusion processes," *Stochastic Analysis and Applications*, vol. 22, no. 6, pp. 1553–1576, 2004.

[50] A. Warshel, "Calculations of chemical processes in solutions," *The Journal of Physical Chemistry*, vol. 83, no. 12, pp. 1640–1652, 1979.

[51] G. G. Batrouni and E. Dagotto, "Accelerated simulations of XY spin glasses," *Phys. Rev. B*, vol. 37, pp. 9875–9878, 1988.

[52] E. Marinari and G. Parisi, "Simulated tempering : A new monte carlo scheme," *Europhysics Letters (EPL)*, vol. 19, no. 6, pp. 451–458, 1992.

[53] W. E, W. Ren, and E. Vanden-Eijnden, "String method for the study of rare events," *Phys. Rev. B*, vol. 66, p. 052301, 2002.

[54] L. Maragliano and E. Vanden-Eijnden, "A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations," *Chemical Physics Letters*, vol. 426, no. 1, pp. 168–175, 2006.

[55] S. Kirkpatrick, J. Gelatt, C.D., and M. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, p. 671, 1983.

[56] V. Černý, "Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm," *Journal of Optimization Theory and Applications*, vol. 45, no. 1, pp. 41–51, 1985.

[57] H. D. Alvarenga, T. V. De Putte, N. Van Steenberge, J. Sietsma, and H. Terryn, "Influence of carbide morphology and microstructure on the kinetics of superficial decarburization of c - mn steels," *Metallurgical and Materials Transactions A*, vol. 46, no. 1, pp. 123–133, 2015.

[58] R. Bellio, S. Ceschia, L. Di Gaspero, A. Schaerf, and T. Urli, "Feature-based tuning of simulated annealing applied to the curriculum-based course timetabling problem," *Computers & Operations Research*, vol. 65, pp. 83–92, 2016.

[59] W. Kerler and P. Rehberg, "Simulated-tempering procedure for spin-glass simulations," *Physical Review E*, vol. 50, no. 5, pp. 4220–4225, 1994.

[60] A. Kone and D. A. Kofke, "Selection of temperature intervals for parallel-tempering simulations," *The Journal of Chemical Physics*, vol. 122, no. 20, p. 206101, 2005.

[61] S. Park and V. S. Pande, "Choosing weights for simulated tempering," *Physical Review E*, vol. 76, no. 1, p. 016703, 2007.

[62] C. Zhang and J. Ma, "Comparison of sampling efficiency between simulated tempering and replica exchange," *The Journal of Chemical Physics*, vol. 129, no. 13, p. 134112, 2008.

[63] G. Behrens, N. Friel, and M. Hurn, "Tuning tempered transitions," *Statistics and Computing*, vol. 22, no. 1, pp. 65–78, 2012.

[64] P. H. Nguyen, Y. Okamoto, and P. Derreumaux, "Communication: Simulated tempering with fast on-the-fly weight determination," *The Journal of Chemical Physics*, vol. 138, no. 6, p. 061102, 2013.

[65] F. Wang and D. P. Landau, "Efficient, multiple - range random walk algorithm to calculate the density of states," *Physical Review Letters*, vol. 86, no. 10, pp. 2050–2053, 2001.

[66] J. Comer, J. C. Gumbart, J. Hénin, T. Lelièvre, A. Pohorille, and C. Chipot, "The adaptive biasing force method : Everything you always wanted to know but were afraid to ask," *The Journal of Physical Chemistry B*, vol. 119, no. 3, pp. 1129–1151, 2015.

[67] T. Zhang, P. H. Nguyen, J. Nasica-Labouze, Y. Mu, and P. Derreumaux, "Folding atomistic proteins in explicit solvent using simulated tempering," *The Journal of Physical Chemistry B*, vol. 119, no. 23, pp. 6941–6951, 2015.

[68] R. H. Swendsen and J.-S. Wang, "Replica monte carlo simulation of spin - glasses," *Physical Review Letters*, vol. 57, no. 21, pp. 2607–2609, 1986.

[69] Y. Sugita and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding," *Chemical Physics Letters*, vol. 314, no. 1-2, pp. 141–151, 1999.

[70] J. Lu and E. Vanden-Eijnden, "Infinite swapping replica exchange molecular dynamics leads to a simple simulation patch using mixture potentials," *The Journal of Chemical Physics*, vol. 138, no. 8, p. 084105, 2013.

[71] K. Nomura, M. Oikawa, A. Kawai, T. Narumi, and K. Yasuoka, "GPU - accelerated replica exchange molecular simulation on solidliquid phase transition study of Lennard-Jones fluids," *Molecular Simulation*, vol. 41, no. 10-12, pp. 874–880, 2015.

[72] T.-Q. Yu, J. Lu, C. F. Abrams, and E. Vanden-Eijnden, "Multiscale implementation of infinite-swap replica exchange molecular dynamics," *Proceedings of the National Academy of Sciences*, vol. 113, no. 42, pp. 11744–11749, 2016.

[73] D. A. Kofke, "On the acceptance probability of replica-exchange monte carlo trials," *The Journal of Chemical Physics*, vol. 117, no. 15, pp. 6911–6914, 2002.

[74] N. Rathore, M. Chopra, and J. J. de Pablo, "Optimal allocation of replicas in parallel tempering simulations," *The Journal of Chemical Physics*, vol. 122, no. 2, p. 024111, 2004.

[75] D. Sindhikara, Y. Meng, and A. E. Roitberg, "Exchange frequency in replica exchange molecular dynamics," *The Journal of Chemical Physics*, vol. 128, no. 2, p. 024103, 2008.

[76] M. D. Donsker and S. Varadhan, "Asymptotic evaluation of certain Markov process expectations for large time, I," *Comm. Pure Appl. Math.*, vol. 28, pp. 1–47, 1975.

[77] J.-D. Deuschel and D. W. Stroock, *Large deviations*. No. v. 137 in Pure and Applied Mathematics, Boston: Academic Press, 1989.

[78] A. Dembo and Ó. Zaitûnî, *Large Deviations Techniques and Applications*. No. 38 in Stochastic Modelling and Applied Probability, Berlin Heidelberg: Springer, 2. ed., corr. print. of the 1998 ed ed., 2010.

[79] J. Lu and E. Vanden-Eijnden, "Methodological and computational aspects of parallel tempering methods in the infinite swapping limit," *Journal of Statistical Physics*, vol. 174, no. 3, pp. 715–733, 2019.

[80] N. Plattner, J. D. Doll, P. Dupuis, H. Wang, Y. Liu, and J. E. Gubernatis, "An infinite swapping approach to the rare-event sampling problem," *The Journal of Chemical Physics*, vol. 135, no. 13, p. 134111, 2011.

[81] G. Gobbo and B. J. Leimkuhler, "Extended hamiltonian approach to continuous tempering," *Phys. Rev. E*, vol. 91, p. 061301, 2015.

[82] Y. Q. Gao, "An integrate-over-temperature approach for enhanced sampling," *The Journal of Chemical Physics*, vol. 128, no. 6, p. 064105, 2008.

[83] G. A. Pavliotis and A. M. Stuart, *Multiscale Methods: Averaging and Homogenization*. No. 53 in Texts in Applied Mathematics, New York: Springer, 2008.

[84] D. Carlson, P. Stinson, A. Pakman, and L. Paninski, "Partition functions from rao-blackwellized tempered sampling," *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48, pp. 2896–2905, 2016.

[85] B. Leimkuhler and C. Matthews, "Rational construction of stochastic numerical methods for molecular sampling," *Applied Mathematics Research eXpress*, vol. 2013, no. 1, pp. 34–56, 2013.

[86] E. Barth, K. Kuczera, B. Leimkuhler, and R. D. Skeel, "Algorithms for constrained molecular dynamics," *Journal of Computational Chemistry*, vol. 16, no. 10, pp. 1192–1209, 1995.

[87] B. Cheng, E. A. Engel, J. Behler, C. Dellago, and M. Ceriotti, "Ab initio thermodynamics of liquid and solid water," *Proceedings of the National Academy of Sciences*, vol. 116, no. 4, pp. 1110–1115, 2019.

[88] K. Yang, X. Xu, B. Yang, B. Cook, H. Ramos, N. M. A. Krishnan, M. M. Smedskjaer, C. Hoover, and M. Bauchy, "Predicting the young s modulus of silicate glasses using high - throughput molecular dynamics simulations and machine learning," *Scientific Reports*, vol. 9, no. 1, pp. 1–11, 2019.

[89] S. Li and Y. Wang, "Percolation phase transition from ionic liquids to ionic liquid crystals," *Scientific Reports*, vol. 9, no. 1, pp. 1–8, 2019.

[90] P. Robustelli, S. Piana, and D. E. Shaw, "Developing a molecular dynamics force field for both folded and disordered protein states," *Proceedings of the National Academy of Sciences*, vol. 115, no. 21, pp. 4758–4766, 2018.

[91] S. Plimpton, "Fast parallel algorithms for short - range molecular dynamics," *Journal of Computational Physics*, vol. 117, no. 1, pp. 1–19, 1995.

[92] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kal, and K. Schulten, "Scalable molecular dynamics with namd," *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1781–1802, 2005.

[93] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen, "Gromacs : A message-passing parallel molecular dynamics implementation," *Computer Physics Communications*, vol. 91, no. 1, pp. 43–56, 1995.

[94] S. E. Feller, Y. Zhang, R. W. Pastor, and B. R. Brooks, "Constant pressure molecular dynamics simulation: The langevin piston method," *The Journal of Chemical Physics*, vol. 103, no. 11, pp. 4613–4621, 1995.

[95] G. Kalibeava, M. Ferrario, and G. Ciccotti, "Constant pressure-constant temperature molecular dynamics: a correct constrained npt ensemble using the molecular virial," *Molecular Physics*, vol. 101, no. 6, pp. 765–778, 2003.

[96] M. D. Pierro, R. Elber, and B. Leimkuhler, "A stochastic algorithm for the isobaric isothermal ensemble with ewald summations for all long range forces," *J. Chem. Theory Comput.*, vol. I, 2015.

[97] H. C. Andersen, "Molecular dynamics simulations at constant pressure and/or temperature," *The Journal of Chemical Physics*, vol. 72, no. 4, pp. 2384–2393, 1980.

[98] G. J. Martyna, D. J. Tobias, and M. L. Klein, "Constant pressure molecular dynamics algorithms," *The Journal of Chemical Physics*, vol. 101, no. 5, pp. 4177–4189, 1994.

[99] L. Simine and P. J. Rossky, "Relating chromophoric and structural disorder in conjugated polymers," *The Journal of Physical Chemistry Letters*, vol. 8, no. 8, pp. 1752–1756, 2017.

[100] W. Gonalves, J. Morthomas, P. Chantrenne, M. Perez, G. Foray, and C. L. Martin, "Elasticity and strength of silica aerogels: A molecular dynamics study on large volumes," *Acta Materialia*, vol. 145, pp. 165 – 174, 2018.

[101] B. Leimkuhler, E. Noorizadeh, and O. Penrose, "Comparing the efficiencies of stochastic isothermal molecular dynamics methods," *Journal of Statistical Physics*, vol. 143, no. 5, pp. 921–942, 2011.

[102] B. Leimkuhler and C. Matthews, *Molecular Dynamics With Deterministic and Stochastic Numerical Methods*. Springer International Publishing, 1 ed., 2015.

[103] K.-i. Okazaki, D. Wöhlert, J. Warnau, H. Jung, Ö. Yildiz, W. Kühlbrandt, and G. Hummer, "Mechanism of the electroneutral sodium/proton antiporter panhap from transition-path shooting," *Nature Communications*, vol. 10, no. 1, pp. 1–10, 2019.

[104] X. Gao, J. Fang, and H. Wang, "Sampling the isothermal-isobaric ensemble by langevin dynamics," *Journal of Chemical Physics*, vol. 144, no. 12, 2016.

[105] N. Grønbech-Jensen and O. Farago, "Constant pressure and temperature discrete-time Langevin molecular dynamics," *The Journal of Chemical Physics*, vol. 141, no. 19, p. 194108, 2014.

[106] S. Cajahuaringa and A. Antonelli, "Stochastic sampling of the isothermal-isobaric ensemble: Phase diagram of crystalline solids from molecular dynamics simulation," *The Journal of Chemical Physics*, vol. 149, no. 6, p. 064114, 2018.

[107] B. Leimkuhler and C. Matthews, "Efficient molecular dynamics using geodesic integration and solvent solute splitting," *Proc. R. Soc. A*, vol. 472, 2016.

[108] M. Parrinello and A. Rahman, "Crystal structure and pair potentials : A molecular - dynamics study," *Physical Review Letters*, vol. 45, no. 14, pp. 1196–1199, 1980.

[109] N. Bou-Rabee and H. Owhadi, "Long- run accuracy of variational integrators in the stochastic context," *SIAM Journal on Numerical Analysis*, vol. 48, no. 1, pp. 278–297, 2010.

[110] E. Marinari and G. Parisi, "Simulated tempering: A new monte carlo scheme," *Europhysics Letters (EPL)*, vol. 19, no. 6, pp. 451–458, 1992.

[111] P. H. Nguyen, Y. Okamoto, and P. Derreumaux, "Communication: Simulated tempering with fast on-the-fly weight determination," *The Journal of Chemical Physics*, vol. 138, no. 6, p. 061102, 2013.

[112] A. BenNaim, "Statistical mechanics of waterlike particles in two dimensions. i . physical model and application of the percusyevick equation," *The Journal of Chemical Physics*, vol. 54, no. 9, pp. 3682–3695, 1971.

[113] R. L. Davidchack, R. Handel, and M. V. Tretyakov, "Langevin thermostat for rigid body dynamics," *The Journal of Chemical Physics*, vol. 130, no. 23, p. 234101, 2009.

[114] R. S. Singh and B. Bagchi, "Solid-solid collapse transition in a two dimensional model molecular system," *The Journal of Chemical Physics*, vol. 139, no. 19, p. 194702, 2013.

[115] A. Scukins, V. Bardik, E. Pavlov, and D. Nerukh, "Molecular dynamics implementation of bn 2d or mercedes benz water model," *Computer Physics Communications*, vol. 190, pp. 129–138, 2015.

[116] T. Urbic, "Liquid part of the phase diagram and percolation line for two-dimensional mercedes - benz water," *Physical Review E*, vol. 96, no. 3, p. 032122, 2017.

[117] S. Auer and D. Frenkel, "Quantitative prediction of crystal–nucleation rates for spherical colloids," *Annual Review of Physical Chemistry*, vol. 55, no. 1, pp. 333–361, 2004.

[118] M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, "Determination of reaction coordinates via locally scaled diffusion map," *The Journal of Chemical Physics*, vol. 134, no. 12, p. 124116, 2011.

[119] M. A. Rohrdanz, W. Zheng, and C. Clementi, "Discovering mountain passes via torchlight : Methods for the definition of reaction coordinates and pathways in complex macromolecular reactions," *Annual Review of Physical Chemistry*, vol. 64, no. 1, pp. 295–316, 2013.

[120] Z. Trstanova, B. Leimkuhler, and T. Lelièvre, "Local and global perspectives on diffusion maps in the analysis of molecular systems," *arXiv:1901.06936*, 2019.