# Bounds on Errors in Observables Computed from Molecular Dynamics Simulations

Vikram Sundar
vikramsundar@college.harvard.edu
Advisor: Dr. David Gelbwaser and Prof. Alán Aspuru-Guzik (Chemistry)
Shadow Advisor: Prof. Martin Nowak

Submitted in partial fulfillment of the honors requirements for the degree of Bachelor of Arts in Mathematics

March 19, 2018

# Contents

# Abstract

Molecular dynamics (MD) is a popular molecular simulation method often used in biochemistry, drug discovery, and materials science. It numerically integrates Newton's second law for a Hamiltonian $H$ determined by a force field, an energy function on the particles in the simulation, and computes condensed phase properties by evaluating a time average of these properties over the simulation. There are three major sources of error in MD simulations: sampling error, due to not running the simulation infinitely long, time discretization error, due to numerical integration of Newton's second law, and force field error, due to modeling a quantum-mechanical molecular system with a classical force field. We demonstrate that under certain nice conditions, sampling and time discretization error are bounded by $\mathcal{O}(T^{-1} + h^2)$, where $T$ is the length of the simulation and $h$ is the step size. Force-field error is more difficult to bound; we examine the assumptions required to derive a force field and provide a rigorous derivation of a bound on nuclear quantum effects, or effects arising from quantum-mechanical properties of the nucleus. We also present an introduction to force-field functor theory, a method of accounting for nuclear quantum effects in force fields. The results of this thesis will help scientists understand when MD simulations can be trusted to provide reliable results.

# Acknowledgments

This thesis would not have been possible without the contributions of a large number of people. Most importantly, I would like to thank Dr. David Gelbwaser and Professor Alán Aspuru-Guzik in the Chemistry Department at Harvard for introducing me to the idea of force-field functors and to the Wigner expansion, and for providing me with invaluable guidance, support, and feedback throughout this project. They helped me develop and present much of the original research in this thesis and guided me through the expository components. I would also like to thank Professor Martin Nowak in the Math Department for serving as my shadow advisor for the course of this project.

This thesis is also the culmination of my intellectual development as an undergraduate who has tried to cross many disciplinary boundaries during my time at Harvard. I have had a number of inspirational professors who have shaped my intellectual perspective. Among many others, I would like to thank Professor Dennis Gaitsgory (math) for teaching me the value of abstraction, Professor Howard Georgi and Dr. Jacob Barandes (physics) for demonstrating the power of physical intuition, Dr. Jacob Sanders (chemistry at UCLA) for explaining the patterns of organic chemistry and connecting them to computer simulation and to physics, and Professor Eric Jacobsen (chemistry) for showing me the power of computational chemistry when used to understand biological and chemical problems. I would also like to thank Tasha Arvanitis, Dr. Alexander Donchev, and the other employees at D. E. Shaw Research for allowing me to intern there for a summer and learn about the field of molecular dynamics.

Finally, I would like to thank my friends at Harvard, especially Melody Tong, for their undying support throughout my undergraduate career, and my parents for seeing the potential in me that I could not and giving me the love and affection I needed to realize it.

# Chapter 1

# Introduction

## 1.1 Molecular Dynamics

### 1.1.1 Why Molecular Dynamics?

One of the fundamental goals of chemistry is to accurately predict the behavior of molecules, whether it be in reactions, binding interactions, solutions, or other forms. Computational chemistry seeks to tackle this challenge by simulating the behavior of these molecules on a computer. Such simulations, if fully accurate, can complement experimental approaches: they can allow scientists to better understand reaction mechanisms, they can help scientists make predictions about molecular behavior without consuming large quantities of expensive chemicals, and they can aid engineers in rational design of materials with particular properties [39].

The focus of this thesis is molecular dynamics (MD), one of the most popular simulation methods due to its relative accuracy and low computational cost [16]. Molecular dynamics has seen application in a wide variety of fields, ranging from drug discovery to material science. It was initially developed as a method to simulate the behavior of liquids and gases of a single molecule, like water or carbon dioxide, and predict their properties. Such simulations were soon expanded to simulations of crystalline substances, of mixtures of liquids, and of solutions [1]. These simulations have helped material scientists predict properties of materials, like conductivity, density, or solubility, and design materials with certain desired properties.

MD simulations have proven even more important in biomedical research. The field of structural biology uses structural changes in biomolecules to understand and make predictions about their function; MD simulations have proven an invaluable complement to experiment in this field to better understand the behavior of large biomolecules. For example, MD simulations have found conformational changes in the shape of proteins that could not be observed in experiments; these conformations were crucial to understanding the behavior and function of those proteins. MD simulations have also been able to model the transport of small molecules through proteins, like those found on the membranes of cells [16].

The most practically useful application of MD simulations is in the prediction of binding between proteins and small molecules, or ligands. MD simulations that model and predict these ligand-protein interactions can then identify drug candidate molecules that would inhibit the activity of drug target proteins. Such simulations have been used in the process of drug discovery, especially since simulations are significantly cheaper and more efficient than the corresponding experimental techniques [16]. MD simulations can also similarly be used to design larger molecules, up to and including proteins, which has

further therapeutic applications in fields like immunotherapy [16].

While MD simulations have a great deal of potential, they, like all computational simulation methods, are hampered by unavoidable errors. MD simulations make a number of approximations, since a totally accurate simulation of a molecular system is at present computationally impractical [1]. Mathematically understanding the errors in MD simulations, and thus errors in observables predicted by MD simulations, is crucial for scientists who use these simulations. Today, many scientists are unwilling to trust the results of MD simulations without independent experimental verification since the errors are often too high [16]. Bounding and understanding these errors would make MD simulations much more powerful and useful [16]. Thus the focus of this thesis will be on bounding the error in observables measured from MD simulations.

### 1.1.2 What is Molecular Dynamics?

#### 1.1.2.1 Force Fields and Integrators

We begin with a brief overview of how the MD simulation algorithm works. MD assumes that all particles in the simulation behave classically, so they move according to Newton's second law

$$\vec{F}(\vec{q}) = m\ddot{\vec{q}} \tag{1.1}$$

where $\vec{F}$ is the force, $\vec{q}$ is the position of our particle, and $m$ is the mass of our particle. Particles in the context of MD simulations generally refer to the nuclei of the atoms. The goal of our simulation is to compute the **trajectory** of our particles, or the path $\vec{q}(t)$ for all particles. We will often also compute the momentum $\vec{p}(t) = m\dot{\vec{q}}(t)$ of all particles [1].

Equation 1.1 is a system of $N$ coupled second-order differential equations given that there are $N$ particles. Obviously, an analytic solution to these differential equations is virtually impossible, so MD simulations will numerically integrate over time in order to compute the trajectory. Numerical integration requires a choice of small time step $h$ and of **integrator** to compute $\vec{q}(nh)$ given $\vec{q}(0), \ldots, \vec{q}((n-1)h)$. The most popular integrator is the Störmer-Verlet method, which we introduce in Chapter 2 [1].

We also need to compute the force $\vec{F}(\vec{q})$ as a function of particle position. This is approximated through a **force field**. Most force fields are actually written in terms of the potential energy $V(\vec{q})$; one may then compute the force by taking $\vec{F}(\vec{q}) = -\vec{\nabla}V(\vec{q})$. These potential functions are typically approximated as additive functions with some of the following terms:

1. Bond stretches between bonded atoms with displacement $\vec{r}$, modeled as harmonic terms

$$V(\vec{r}) = \frac{1}{2}k(\vec{r} - \vec{r}_0)^2$$

   for constant $k, \vec{r}_0$.

2. Bond angles between pairs of bonds with angle $\theta$ between them, modeled as harmonic terms

$$V(\theta) = \frac{1}{2}k(\theta - \theta_0)^2$$

   for constant $k, \theta_0$.

3. Bond torsions between a series of three bonds with dihedral angle $\phi$ between them, modeled as a low-order Fourier expansion

$$V(\phi) = c_0 + c_1 \sin \phi + c_3 \sin 3\phi + c_5 \sin 5\phi$$

   for constant $c_0, c_1, c_3, c_5$.

4. Medium-range van der Waals repulsion terms between any two atoms at distance $r$, modeled as

$$V(r) = \frac{A}{r^{12}} - \frac{B}{r^6}$$

   for constants $A, B$.

5. Long-range electrostatic interactions between any two charged atoms with charges $q_1, q_2$ at distance $r$, modeled by Coulomb's law as

$$V(r) = -\frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r}$$

   with $\epsilon_0$ the vacuum permittivity.

More modern force fields have many more complicated terms, but the basic idea remains the same: we have some sum of some number of terms with constants. These constants are generally fixed by fitting to either experimental or quantum mechanical data [35]. Chapters 5 and 6 provide details about how force fields are typically derived.

Once the force field and integrator are selected, the basic structure of an MD simulation is very simple:

1. Initialize the system (typically to a representative conformation).

2. At each time step, compute the forces using the force field.

3. Use the integrator to compute positions at the next time step.

### 1.1.2.2 Observables

In order to ensure that the MD simulation accurately represents reality, there are a few additional steps that must be taken. First, the initial conditions must be realistic. We typically avoid boundary effects in our system by simulating a torus $\mathbb{T}^3$ instead of a fixed-size box. Practically speaking, this simply means that particles exiting one side of the box will enter the opposite side. Selecting initial conditions beyond the periodic box is done on a case-by-case basis [1].

Second, we note that the simulation algorithm outlined above automatically conserves energy and volume, since all forces are defined in terms of a potential and we are simulating particles inside a box of fixed size. This algorithm represents an **NVE simulation**, since it conserves $N$, particle number, $V$, the volume of the box, and $E$, the total energy. Experimental conditions under which measurements are taken rarely conserve these quantities; they usually conserve temperature and pressure instead. Thus we would

prefer to run an **NVT simulation** (conserving temperature $T$ instead of energy $E$) or an **NPT simulation** (conserving pressure $P$ instead of volume $V$).

We do this by using Monte Carlo moves. Specifically, if we want to conserve temperature $T$, after each stage, we run a brief Monte Carlo simulation step to perturb all the particle's energies; if we want to conserve pressure $P$, we perturb the box's volume. The probabilities of these moves being accepted depends on the desired temperature $T$ or pressure $P$. This method allows us to essentially reduce the NPT or NVT simulation to an NVE simulation; further, since Monte Carlo methods are well-known to converge in the long-time limit, this method adds essentially no error to our analysis. Thus we will focus on NVE simulations for most of this thesis, knowing that our results will also apply for the more experimentally useful NVT or NPT simulations [1].

We may now compute an observable $A$. For this thesis, we will assume that $A$ is a function of the position $\vec{q}$ and momentum $\vec{p}$ of all the particles. Thus our simulation, after running for a time $T = nh$, will compute our observable by averaging over all frames in our trajectory

$$\langle A \rangle^{\text{Rie}}(T) = \frac{1}{n} \sum_{i=0}^{n} A(\vec{q}(ih), \vec{p}(ih)). \tag{1.2}$$

Note that our trajectory here is determined by simulation on an approximate force field. Our experiment will measure this observable also by averaging, but experiments generally run for much longer than simulations do and thus explore many many more configurations. We generally assume that experiments explore essentially all accessible configuration space, where accessible means that quantities that must be conserved (like energy) are conserved. This assumption is valid since experiments involve a large number of particles in a macroscopic system that are randomly distributed over all accessible configuration space; we call such averages **ensemble averages**. So we have

$$\langle A \rangle = \int_M A(\vec{p}, \vec{q}) \, d\vec{p} \, d\vec{q} \tag{1.3}$$

where $M$ is the manifold corresponding to the accessible configuration space that we integrate over. Here, our manifold $M$ is determined by experiments run on the actual physical system, not the approximate force field. Physical intuition leads us to believe that $\langle A \rangle^{\text{Rie}}(T)$ and $\langle A \rangle$ are close to each other since a randomly initialized system over long periods of time $T$ should explore all of configuration space. We will define $\langle A \rangle^{\text{Rie}}(T)$ and $\langle A \rangle$ more precisely and gain a better understanding of the difference between them in this thesis.

The fundamental question we seek to answer in this thesis is the following: **Are the quantities $\langle A \rangle^{\text{Rie}}(T)$ and $\langle A \rangle$ as defined above close to each other? Can we derive bounds on how close they are to each other?**

## 1.2 Structure of this Thesis

### 1.2.1 Sources of Error

Broadly speaking, there are three major sources of error that are found in MD simulations and contribute to the difference between $\langle A \rangle^{\text{Rie}}(T)$ and $\langle A \rangle$ [4]:

1. **Sampling error**. This results from the fact that we truncate our simulation at some finite time $T$ instead of letting our simulation run until $T = \infty$. The size of sampling error depends on our integrator and the length of the simulation $T$.

2. **Time discretization error**. This results from the fact that we choose to numerically integrate Newton's second law with some finite time step $h$ instead of analytically solving Newton's second law with an infinitesimal time step. The size of time discretization error depends on our integrator and the size of the time step $h$.

3. **Force field error**. This error refers to any error that is a result of inaccuracies in our force field, not inaccuracies in our integrator. Fundamentally, it results from the fact that molecular systems are quantum mechanical and MD simulations model them classically, which requires that some approximations be made in order to consider quantum effects.

The goal of this thesis is to present mathematically proven bounds on these errors. Bounding these errors in general in an MD simulation is still an open problem, so the results we present here only cover a small number of cases that comply with some simplifying conditions. We also provide an exposition to what progress needs to be made to understand these error terms and which errors are considered to be more or less important for accurate simulations.

## 1.2.2 Key Results of this Thesis

In Chapter 2, we present important properties of the primary integrator of an MD simulation, the Störmer-Verlet method, and show that its properties guarantee that energy is conserved over polynomial times in Theorem 2.2.3. This result, while not a bound on the error terms listed above, is still important since it reassures us that energy is conserved in an NVE simulation, as expected.

In Chapter 3, we assume that our system is completely integrable, derive important properties of completely integrable systems including the Arnold-Liouville Theorem (Theorem 3.1.9), and explain why MD simulations are likely to be close to completely integrable. We then use these results to bound sampling error to $\mathcal{O}(T^{-1})$ in Theorem 3.2.2 and demonstrate that the use of a filter function allows sampling error to be bounded by an exponentially decaying function in $T$ in Theorem 3.2.3.

In Chapter 4, we introduce perturbation theory and use it to derive bounds on time discretization error. The key result here is a $\mathcal{O}(T^{-1} + h^2)$ bound on time discretization and sampling error in Theorem 4.2.2. This theorem summarizes the error bounds that result from the integrator.

In the last two chapters of this thesis, we focus on force field error. This is a broad field, so we focus on a few key mathematically proven results and on developing some physical intuition behind common sources of force field error. In Chapter 5, we provide a brief overview of quantum mechanics as applied to molecular systems and then describe the Born-Oppenheimer approximation, the key result allowing us to approximate quantum mechanical systems using MD simulations with classical force fields. In Section 5.2.2, we qualitatively describe a number of the sources of error associated with Born-Oppenheimer force fields.

The final sections of this thesis focus on nuclear quantum effects, the subject of my own research. While these are not necessarily the largest source of error in force field development, results about nuclear quantum effects affect almost every force field developed. We begin this discussion by proving a complicated bound on nuclear quantum effects in Born-Oppenheimer force fields in Theorem 5.2.1; this bound makes clear that as long as nuclei have finite mass, nuclear quantum effects will contribute an error term that does not vanish. In Chapter 6, we introduce force-field functor theory, a way of accounting for nuclear quantum effects in Born-Oppenheimer force fields. Theorem 6.1.1 tells us that it is always possible to derive an effective force field that accounts for nuclear quantum effects and Section 6.2.1 provides an efficient means of computing this effective force field along with empirical justification of its effectiveness.

# Chapter 2

# The Störmer-Verlet Method and Energy Conservation

We begin by introducing the primary integrator used in MD simulations, the **Störmer-Verlet method**. Newton's second law from Equation 1.1 is a second order differential equation of the form

$$\ddot{\vec{q}} = f(\vec{q})$$

for some function $f$ (the force divided by mass) and some position-space vector $\vec{q}$ with dimension equal to the number of degrees of freedom. For simplicity, focus on the one-dimensional case $\ddot{q} = f(q)$ for now. Introduce the **velocity** $\dot{q} = v$, so our equation is a coupled system

$$\dot{v} = f(q)$$
$$\dot{q} = v.$$

The space defined by $(v, q)$ (or the corresponding space defined by $(p, q)$ for $p = mv$) is known as the **phase space** [25].

Now discretize our system with timepoints $t_n = nh$ (for some small step size $h$ and $n$ a positive integer) and corresponding positions $q_n$ and velocities $v_n$ (here the indexing corresponds to timesteps). We may now evolve $\Phi : (q_n, v_n) \mapsto (q_{n+1}, v_{n+1})$ by the following three steps:

$$v_{n+\frac{1}{2}} = v_n + \frac{h}{2} f(q_n)$$
$$q_{n+1} = q_n + h v_{n+\frac{1}{2}} \tag{2.1}$$
$$v_{n+1} = v_{n+\frac{1}{2}} + \frac{h}{2} f(q_{n+1}).$$

See Figure 2.0.1a. Here, we introduce staggered timepoints $t_{n+\frac{1}{2}} = \left(n + \frac{1}{2}\right) h$. There is a dual option to Equation 2.1 where we switch the roles of $v$ and $q$:

$$q_n = q_{n-\frac{1}{2}} + \frac{h}{2} v_{n-\frac{1}{2}}$$
$$v_{n+\frac{1}{2}} = v_{n-\frac{1}{2}} + h f(q_n) \tag{2.2}$$
$$q_{n+\frac{1}{2}} = q_n + \frac{h}{2} v_{n+\frac{1}{2}}$$

See Figure 2.0.1b. Now observe that we may simply concatenate the last and first steps of both methods in Equations 2.1 and 2.2 together. This gets us the Störmer-Verlet method.

(a) One Symplectic
Integrator.

(b) Its Dual.
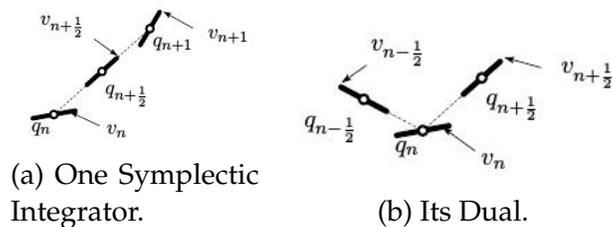
Figure 2.0.1: Motivation for Störmer-Verlet. Taken from Hairer et al. [25].

**Definition 2.0.1.** The **Störmer-Verlet method** is the map $\Phi_h : (q_n, v_{n-\frac{1}{2}}) \mapsto (q_{n+1}, v_{n+\frac{1}{2}})$ such that:

$$v_{n+\frac{1}{2}} = v_{n-\frac{1}{2}} + hf(q_n)$$
$$q_{n+1} = q_n + hv_{n+\frac{1}{2}}$$

(2.3)

The formula above can be trivially generalized to multiple dimensions.

This is the primary integrator of choice for MD simulations, both because of its ease of implementation and its multitude of nice properties that allow reasonably fast convergence of these simulations. Our goal for this chapter will be to show that the Störmer-Verlet method belongs to a class of integrators called symplectic integrators, which conserve energy over long timescales [25]. For a comparison of this method with others, see Hairer et al. [26], Arnold [2].

## 2.1 Symplecticity

### 2.1.1 Hamiltonian and Lagrangian Mechanics

Our system of differential equations is a special one determined by Newton's second law. Thus we have some important additional properties, deriving from the Lagrangian and Hamiltonian formalism of classical mechanics.

We may write

$$K = \frac{1}{2}\dot{\vec{q}} \cdot (M(\vec{q})\dot{\vec{q}})$$

for $K$ the kinetic energy and $M(q)$ the mass matrix, which will always be symmetric and positive-definite. We are also given a force field, a potential energy function $U = U(q)$. The **Lagrangian** is then

$$L = K - U$$

and the Euler-Lagrange equations stipulate

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\vec{q}}}\right) = \frac{\partial L}{\partial \vec{q}}$$

(2.4)

which in our case imply Newton's second law in Equation 1.1.

We may now define the **conjugate momenta**

$$\vec{p} = \frac{\partial L}{\partial \dot{\vec{q}}}(\vec{q}, \dot{\vec{q}}) = M(\vec{q})\dot{\vec{q}}$$

for all $i$ and the **Hamiltonian**

$$H(\vec{p}, \vec{q}) = \vec{p} \cdot \dot{\vec{q}} - L(\vec{q}, \dot{\vec{q}}) = \frac{1}{2} \vec{p} \cdot (M(\vec{q})^{-1} \vec{p}) + U(\vec{q}).$$

This is the total energy of our system, and it is well-known that it is conserved on any trajectory [26]. Further:

**Theorem 2.1.1.** *The Euler-Lagrange equations in Equation 2.4 are equivalent to the Hamiltonian equations of motion*

$$\dot{p}_k = -\frac{\partial H}{\partial q_k}(p, q), \dot{q}_k = \frac{\partial H}{\partial p_k}(p, q). \tag{2.5}$$

*Proof from Arnold [2].* Observe that

$$\frac{\partial H}{\partial q_k} = \frac{\partial}{\partial q_k}(\vec{p} \cdot \dot{\vec{q}} - L(\vec{q}, \dot{\vec{q}})) = -\frac{\partial L}{\partial q_k}$$

and similarly $\frac{\partial H}{\partial p_k} = \dot{q}_k$. The second is directly one of Hamilton's equations of motion; the first simply requires applying the definition of momentum.

For the other direction,

$$\frac{\partial L}{\partial q_k} = \frac{\partial}{\partial q_k}(\vec{p} \cdot \dot{\vec{q}} - H(\vec{p}, \vec{q})) = \dot{\vec{p}} = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\vec{q}}}\right)$$

as desired. □

### 2.1.2 Symplectic Structures on Manifolds

Flows in phase space along the Hamiltonian equations of motion satisfy an additional important property known as symplecticity. Over the next few sections, we will see that symplecticity corresponds to preserving areas in phase space under Hamiltonian flows.

Symplecticity can be defined on any $2d$-dimensional manifold $M$. For future convenience and to simplify proofs of a number of important results, I will define symplecticity in complete generality and avoid specializing to $\mathbb{R}^{2d}$ until absolutely necessary.

**Definition 2.1.2.** A **symplectic structure** on $M$ is a closed nondegenerate differential 2-form $\omega$, i.e. $d\omega = 0$ and for all $\vec{v} \neq \vec{0}$ there exists $\vec{w}$ such that $\omega(\vec{v}, \vec{w}) \neq 0$.

**Example 2.1.3.** The canonical example of a symplectic structure, and the one we will be most interested in, is the following. Consider phase space $\mathbb{R}^{2d}$ with coordinates $p_i, q_i$. Then $\omega = \sum_i dp_i \wedge dq_i$ is symplectic. This can be easily verified directly.

We can understand the above example more generally. Consider $V$, an $d$-dimensional manifold. It is well-known that $T^*V$, the cotangent bundle, is a $2d$-dimensional differentiable manifold. If we choose coordinates $q_i$ for points in $V$, we can choose corresponding coordinates $p_i$ for any cotangent vector in $T_x^*V$ given a point $x \in V$, and thus have coordinates $p_i, q_i$ for all of $T^*V$. Thus the symplectic structure $\omega$ defined above can be naturally extended to $T^*V$ [2].

Now fix a point $x \in V$. We may define a map $I : T_x M \to T_x^* M$ by sending $\vec{v} \mapsto \iota_{\vec{v}}(\omega) = \omega_{\vec{v}}$ where $\omega_{\vec{v}}(\vec{w}) := \omega(\vec{v}, \vec{w})$; this corresponds to the interior product. $I$ is an isomorphism because symplectic forms are nondegenerate. Thus given a Hamiltonian function $H : M \to \mathbb{R}$, we have a differential 1-form $dH$ and a corresponding vector field $\overline{H}$ associated with it, known as the **Hamiltonian vector field** [2].

**Example 2.1.4.** Let's calculate what this isomorphism looks like in $\mathbb{R}^{2d}$. Given a basis vector $e_{p_i}$ it maps to $\omega_{p_i}(\vec{w}) := \omega(\vec{e}_{p_i}, \vec{w}) = dq_i(\vec{w})$ thus we have $\omega_{p_i} = dq_i$. Similarly, $\omega_{q_i}(\vec{w}) := \omega(\vec{e}_{q_i}, \vec{w}) = -dp_i(\vec{w})$ thus we have $\omega_{q_i} = -dp_i$. The matrix of the appropriate transformation in the $p_i, q_i$ basis is $\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$ where $I$ is the $d$-dimensional identity matrix.

Now suppose we have a Hamiltonian $H$. Then

$$\overline{H} = \sum_i \vec{e}_{p_i} \frac{\partial H}{\partial q_i} - \sum_i \vec{e}_{q_i} \frac{\partial H}{\partial p_i}$$

so the corresponding flow equation $\dot{x} = \overline{H}(x)$ allows us to re-derive Hamiltonian dynamics.

Thus our notion of Hamiltonian dynamics agrees with Hamiltonian vector fields (hence the name).

### 2.1.3 Hamiltonian Flows are Symplectic

Given a symplectic manifold $M$ and a Hamiltonian function $H$, the corresponding Hamiltonian vector field $\overline{H}$ generates a **Hamiltonian phase flow** $\varphi_t : M \to M$ such that

$$\varphi_0(x) = x, \frac{d}{dt}\varphi_t(x) = \overline{H}(\varphi_t(x)).$$

This phase flow simply tracks the evolution of our system in phase space. The importance of symplecticity is that:

**Theorem 2.1.5** (Poincaré). *The Hamiltonian phase flow is **symplectic**, or preserves the symplectic structure of the manifold, i.e. $(\varphi_t)^* \omega = \omega$.*

If we work in 1 dimension with $M = \mathbb{R}^2$, this theorem implies that the phase flow preserves the area in phase space. In higher dimensions with $M = \mathbb{R}^{2n}$, this theorem is equivalent to preserving the area under projection to the $(p_i, q_i)$ plane, i.e. the phase space for each individual coordinate.

*Proof from Arnold [2].* Observe that $\varphi_t$ defines a homotopy trivially. Given a manifold $M$ and a $k$-chain $c$ on it, define a $(k+1)$-chain $\varphi c$, the track of chain $c$ under the homotopy $\varphi_t$ for $t \in [0, \tau]$. For each cell in this chain $f : D \to M$, take the map $g : D \times [0, \tau] \to M$ defined by $g(x, t) = \varphi_t(f(x))$ and add the appropriate orientation, so $\varphi c$ is swept out by $c$ under the homotopy $\varphi_t$ for $t \in [0, \tau]$. We may compute

$$\partial(\varphi c) = \varphi_\tau c - c - \varphi \partial c.$$

**Lemma 2.1.6.** *Given a 1-chain c, we have*

$$\frac{d}{dt} \int_{\varphi c} \omega = \int_{\varphi_\tau c} dH.$$

*Proof.* Pick a chain with one cell $f : [0,1] \to M$. Let $g(s,t) = \varphi_t f(s), \vec{v} = \frac{\partial g}{\partial s}, \vec{w} = \frac{\partial g}{\partial t}$ in the tangent space $T_{g(s,t)} M$. Then

$$\int_{\varphi c} \omega = \int_0^1 \int_0^\tau \omega(\vec{v}, \vec{w}) \, dt \, ds$$

but by the definition of Hamiltonian phase flow we have $\vec{w} = \overline{H}(g(s,t))$ so $\omega(\vec{v}, \vec{w}) = dH(v)$. Thus

$$\int_{\varphi c} \omega = \int_0^\tau \left( \int_{\varphi_\tau c} dH \right) dt.$$

We can now differentiate to get the result. $\qquad\square$

Now given any 2-chain $c$, we have

$$0 = \int_{\varphi c} d\omega = \int_{\partial \varphi c} \omega = \int_{\varphi_\tau c} \omega - \int_c \omega - \int_{\varphi \partial c} \omega$$

by Stokes's theorem. We know that

$$\int_{\varphi \partial c} \omega = \int_0^\tau \left( \int_{\varphi_\tau \partial c} dH \right) dt = \int_0^\tau \left( \int_{\partial \varphi_\tau \partial c} H \right) dt = 0$$

by Lemma 2.1.6 so we get

$$\int_{\varphi_\tau c} \omega - \int_c \omega = 0.$$

Since this must be true on any chain, we have $(\varphi_t)^* \omega = \omega$. $\qquad\square$

Let's consider the reverse condition: are flows that preserve the symplectic structure Hamiltonian? This isn't quite always true; consider the torus $T^2$ with coordinates $(p,q)$ and $\omega = dp \wedge dq$. Consider $\varphi_t(p,q) = (p+t,q)$; this clearly preserves the symplectic structure. We can see that we require $H = -q + C$ but $q$ is only a local coordinate, so this can't hold globally.

However, we can prove something local.

**Definition 2.1.7.** A **locally Hamiltonian vector field** is a vector field $\overline{\alpha}$ corresponding under $I$ to a closed 1-form $\alpha$.

Locally, we may write $\alpha = dH$ by Poincaré's lemma and we have a local Hamiltonian function, but this may not hold globally. Then we have:

**Theorem 2.1.8.** *A flow $\varphi_t$ preserves the symplectic structure iff it is locally Hamiltonian.*

*Proof from Pelayo [46].* The backwards direction is immediate from Poincaré's theorem, Theorem 2.1.5, since the proof of that was entirely local.

For the forwards direction, suppose we have such a flow $\varphi_t$. Now consider the vector field $\frac{d}{dt}\varphi_t(x)$. Under $I$ this maps to some $\alpha$ such that $\alpha(v) = \omega\left(\frac{d}{dt}\varphi_t(x), v\right)$. Now preserving the symplectic structure implies that the Lie derivative $\pounds_{\frac{d}{dt}\varphi_t(x)}\omega = 0$, so by the formula $\pounds_v = d \circ \iota_v + \iota_v \circ d$, we get that $\alpha$ is closed, as desired. $\qquad\square$

We can further see that the obstruction to a symplectic flow from being globally Hamiltonian corresponds directly to closed forms that are not exact, i.e. to the first cohomology group $H^1(M, \mathbb{R})$. On any simply connected manifold, i.e. $M = \mathbb{R}^{2d}$, this is just 0, so all flows that preserve the symplectic structure are Hamiltonian [46, 2].

### 2.1.4 Symplecticity of the Störmer-Verlet Method

Since the flow of our Hamiltonian system is symplectic, we would like to verify that the numerically computed flow of our system via the Störmer-Verlet method is also symplectic. As it turns out, this is case:

**Theorem 2.1.9.** *The flow of the Störmer-Verlet method $\Phi_h$ on a Hamiltonian system is symplectic.*

This theorem is true for general Hamiltonians [25], but the elegant proof we present holds only in the case that the Hamiltonian is separable $H(p, q) = K(p) + U(q)$. This is the case for all force fields.

*Proof from Hairer et al. [25].* Let $\varphi_h^K$ be the accurate flow of the Hamiltonian system $H(p, q) = K(p)$ over a time step $h$ and likewise for $\varphi_h^U$. Observe that Equation 2.1 implies that $\Phi_h = \varphi_{\frac{h}{2}}^U \circ \varphi_h^K \circ \varphi_{\frac{h}{2}}^U$. By Theorem 2.1.5, each of the $\varphi$s is symplectic, so so is $\Phi_h$. Since this is equivalent to Störmer-Verlet with some slight index modifications, we have that the flow of the Störmer-Verlet method is also symplectic. $\qquad\square$

## 2.2 Backward Error Analysis

The primary tool for analyzing errors associated with the Störmer-Verlet method and other methods of numerically integrating differential equations is **backward error analysis**. In this section, we will use backward error analysis on the Störmer-Verlet method to derive bounds on long-time energy conservation; later in Chapter 4, we will use it to bound time discretization error.

Suppose we have an exact differential equation $\dot{x} = f(x)$ with solution $\varphi_t(x)$. The numerical solution is $x_n$ with $x_{n+1} = \Phi_h(x_n)$. The idea of backward error analysis is to find a modified differential equation

$$\dot{\tilde{x}} = f_h(\tilde{x}) = f(x) + hf_2(x) + h^2 f_3(x) + \dots \tag{2.6}$$

and determine the difference between $f$ and $f_h$. This gives more insight into the error of the trajectory $x_n$. We can demonstrate existence of this series, but we do not guarantee its convergence.

**Theorem 2.2.1** (Backward Error Analysis Series Existence.). *Consider the equation $\dot{\vec{x}} = f(\vec{x})$ over some vector $\vec{x}$ and infinitely differentiable vector field $f(\vec{x})$. If the numerical method has a Taylor expansion*

$$\Phi_h(\vec{x}) = \vec{x} + hf(\vec{x}) + h^2 g_2(\vec{x}) + h^3 g_3(\vec{x}) + \dots$$

*then there exist unique vector fields $f_i(x)$ such that*

$$\Phi_h(\vec{x}) = \tilde{\varphi}_{h,N}(\vec{x}) + \mathcal{O}(h^{N+1})$$

*and $\tilde{\varphi}_{h,N}$ is the flow of the modified equation*

$$\dot{\vec{x}} = f(\vec{x}) + hf_2(\vec{x}) + \dots + h^{N-1}f_N(\vec{x}). \tag{2.7}$$

This proof is written for the one-dimensional case for simplicity, but the ideas can be easily extended to higher dimensions.

*Proof from Hairer et al. [25].* Define $\tilde{x}(t) = \tilde{\varphi}_t(x)$. We can now expand the flow of Equation 2.6 as a Taylor series to get

$$\tilde{\varphi}_h(x) = x + h\dot{\tilde{x}}(0) + \frac{h^2}{2!}\ddot{\tilde{x}}(0) + \dots$$

$$= x + h\left(f(x) + hf_2(x) + h^2 f_3(x) + \dots\right) + \frac{h^2}{2!}\left(f'(x) + hf_2'(x) + \dots\right)\left(f(x) + hf_2(x) + \dots\right)$$

and comparing like powers with the given equations we get

$$f_2(x) = g_2(x) - f'f(x)$$

and so on. $\square$

Given this proof, observe that if our numerical method is of order $p$, i.e.

$$\Phi_h(\vec{x}) = \varphi_h(\vec{x}) + h^{p+1}\delta_{p+1}(\vec{x}) + \mathcal{O}(h^{p+2})$$

where $\varphi_h(\vec{x})$ is the exact flow and $\delta_{p+1}(x)$ is the leading-order correction, then $f_j(x) = 0$ for $2 \leqslant j \leqslant p$.

Explicit computation for the Störmer-Verlet method [25] gets

$$\Phi_h(\vec{q}, \vec{v}) = \begin{pmatrix} \vec{q} + h\vec{v} + \frac{h^2}{2}f(\vec{q}) \\ \vec{v} + \frac{h}{2}f(\vec{q}) + \frac{h}{2}f\left(\vec{q} + h\vec{v} + \frac{h^2}{2}f(\vec{q})\right) \end{pmatrix}$$

so

$$f_2(\vec{q}, \vec{v}) = 0, \, f_3(\vec{q}, \vec{v}) = \frac{1}{12}\begin{pmatrix} -2f'(\vec{q})\vec{v} \\ f'(\vec{q})f(\vec{q}) + f''(\vec{q})(\vec{v}, \vec{v}) \end{pmatrix}, \, f_4(\vec{q}, \vec{v}) = 0.$$

It can be verified that the Störmer-Verlet method has order 2, so these results agree with what we expect [25].

Note that Störmer-Verlet is symmetric, i.e. $\Phi_h = \Phi_{-h}$. Thus our expansion must also have this property, implying that even powers $f_{2j}(\vec{q}, \vec{v})$ will vanish [25].

### 2.2.1 Symplectic Methods and Backward Error Analysis

Symplectic methods have particularly nice modified differential equations. Specifically, they will also be symplectic and thus also have modified Hamiltonians.

**Theorem 2.2.2.** *A symplectic method $\Phi_h$ applied to a Hamiltonian system with a smooth Hamiltonian H results in a modified equation that is also Hamiltonian, i.e. each truncated equation from Equation 2.7 is also Hamiltonian for some $H_j$.*

*Proof from Hairer et al. [26].* Induct on $j$. The base case is immediate since $f(\vec{x}) = f_1(\vec{x})$.

For the inductive step, note that

$$\dot{\tilde{\vec{x}}} = f(\tilde{\vec{x}}) + hf_2(\tilde{\vec{x}}) + \ldots + h^{r-1}f_r(\tilde{\vec{x}})$$

is Hamiltonian by induction. Let its flow be $\varphi_{r,h}(\vec{x}_0)$. We have

$$\Phi_h(\vec{x}_0) = \varphi_{r,h}(\vec{x}_0) + h^{r+1}f_{r+1}(\vec{x}_0) + \mathcal{O}(h^{r+2})$$

and taking the derivative

$$\Phi'_h(\vec{x}_0) = \varphi'_{r,h}(\vec{x}_0) + h^{r+1}f'_{r+1}(\vec{x}_0) + \mathcal{O}(h^{r+2}).$$

Note that $\Phi_h$ and $\varphi_{r,h}$ are symplectic by induction and using Theorem 2.1.5. Further, $\varphi'_{r,h}(\vec{x}_0) = I + \mathcal{O}(h)$ by definition. We claim that $f_{r+1}$ is symplectic, i.e. $\omega(f'_{r+1}\vec{v}, f'_{r+1}\vec{w}) = \omega(\vec{v}, \vec{w})$. This is immediate by expanding the equation above and taking terms to the $h^{r+1}$ order. Now, apply Theorem 2.1.8 to get the desired. $\square$

### 2.2.2 Bounds on Energy Conservation

This set-up is sufficient for us to prove a significant result on how closely energy is conserved under the Störmer-Verlet method.

**Theorem 2.2.3.** *Given fixed step size h and number of steps n, let the time $t = nh$. Then there exist constants C and $C_N$ independent of t and h such that*

$$|H(\vec{p}_n, \vec{q}_n) - H(\vec{p}_0, \vec{q}_0)| \leqslant Ch^2 + C_N h^N t$$

*if $0 \leqslant t \leqslant h^{-N}$ for some positive integer N. $C_N$ depends on up to the $(N + 1)$th order derivative of H in a region containing $(\vec{p}_n, \vec{q}_n)$.*

*Proof from Hairer et al. [25].* Truncate our modified differential equation after $N$ terms. Apply Theorem 2.2.2 to get a modified Hamiltonian $\tilde{H}(\vec{p}, \vec{q})$ associated with the Störmer-Verlet method. By the Triangle Inequality, we have

$$|H(\vec{p}_n, \vec{q}_n) - H(\vec{p}_0, \vec{q}_0)| \leqslant |H(\vec{p}_n, \vec{q}_n) - \tilde{H}(\vec{p}_n, \vec{q}_n)| + |\tilde{H}(\vec{p}_n, \vec{q}_n) - \tilde{H}(\vec{p}_0, \vec{q}_0)| + |\tilde{H}(\vec{p}_0, \vec{q}_0) - H(\vec{p}_0, \vec{q}_0)|.$$

The first and third terms are bounded by the error in the modified Hamiltonian. Recall that the Störmer-Verlet method has order 2; this gives us a $\mathcal{O}(h^2)$ error in the proof of Theorem 2.2.2. Thus we have

$$|H(\vec{p}_n, \vec{q}_n) - \tilde{H}(\vec{p}_n, \vec{q}_n)| + |\tilde{H}(\vec{p}_0, \vec{q}_0) - H(\vec{p}_0, \vec{q}_0)| \leqslant Ch^2$$

for some fixed constant $C$.

Rewrite the last sum by the Triangle Inequality as

$$|\tilde{H}(\vec{p}_n, \vec{q}_n) - \tilde{H}(\vec{p}_0, \vec{q}_0)| \leqslant \sum_{i=1}^{n} |\tilde{H}(\vec{p}_i, \vec{q}_i) - \tilde{H}(\vec{p}_{i-1}, \vec{q}_{i-1})|.$$

Our modified flow $\tilde{\phi}_t$ preserves the modified Hamiltonian, so

$$|\tilde{H}(\vec{p}_i, \vec{q}_i) - \tilde{H}(\vec{p}_{i-1}, \vec{q}_{i-1})| = |\tilde{H}(\vec{p}_i, \vec{q}_i) - \tilde{H}(\tilde{\phi}_h(\vec{p}_{i-1}, \vec{q}_{i-1}))| \leqslant C_N h^{N+1}$$

since by construction $\tilde{\phi}_h(\vec{p}_{i-1}, \vec{q}_{i-1}) = (\vec{p}_i, \vec{q}_i) + \mathcal{O}(h^{N+1})$. $C_N$ here is clearly only dependent on up to the $(N+1)$th order derivative of $H$, since that's all the modified differential equation is dependent on. Adding up all $n$ terms, we get a bound of $C_N h^N t$, as desired. $\qquad \square$

This is not the sharpest possible bound; you can show that for analytic Hamiltonians $H$ (all Hamiltonians used in molecular dynamics simulations) we have a bound of $Ch^2 + C_0 e^{-\frac{c}{h}t}$ for exponentially long times $t \leqslant e^{\frac{c}{h}}$.

However, the exact bound is unimportant for applications to simulations. The important result of the theorem is that energy is approximately conserved over long time periods. As a result, scientists running NVE MD simulations are guaranteed that energy will actually be conserved over the course of their simulation, so the simulated trajectories will be representative of the real world [1]. Symplecticity of the Störmer-Verlet method is essential for this result to hold; in particular, non-symplectic methods such as the Euler method fail to conserve energy and can lead to nonsensical results in simulation, such as a harmonic oscillator's trajectory becoming totally unbounded [26].

This result provides a sanity check on the expected accuracy of our simulations, but it does not tell us anything about the bounds we expect on errors in observable quantities measured from a simulation. We turn to that subject next, but we'll need a few further assumptions on our system to make progress.

# Chapter 3

# Integrable Systems and Bounds on Sampling Error

The goal of this chapter will be to compute bounds on sampling error, which can be defined as the between the time average

$$\langle A \rangle (T) = \frac{1}{T} \int_0^T A(\vec{p}(t), \vec{q}(t)) \, dt$$

and the spatial average $\langle A \rangle$. We begin with a discussion of the motivation behind our approach and how it differs from some other approaches taken by computational scientists using simulations.

The most natural way to compare time and spatial averages of simulated systems is by citing the **ergodic hypothesis**. The version of the ergodic hypothesis we will use from [18] is occasionally called the quasi-ergodic hypothesis or the modified ergodic hypothesis:

**Conjecture 3.0.1** (Ergodic Hypothesis). *The trajectory of a system is dense on its energy surface as a subset of phase space (the space of all possible positions and momenta of all particles).*

The ergodic hypothesis immediately leads to the conclusion that time averages at infinite time and spatial averages are equal, so time averages should converge to spatial averages. However, on its own, it doesn't tell us any information about the rate of that convergence. Further, it is notoriously difficult to prove that even simple systems are ergodic; no one has attempted to prove whether a full MD force field applied to a protein is ergodic. Even the ergodicity of one of the simplest terms in a force field, the Lennard-Jones term, is an open problem [18].

Known results about ergodic systems make the ergodicity assumption very troubling. First, the space of ergodic Hamiltonians is meager in the space of all Hamiltonians, i.e. covered by a union of nowhere dense subsets [36]. This implies that arbitrarily small perturbations to any ergodic Hamiltonian can result in a non-ergodic Hamiltonian and that we may have non-ergodic Hamiltonians that are far away from ergodic Hamiltonians. Thus given that we know nothing about the ergodicity of our force field, assuming ergodicity does not seem like a good idea.

Further, we have good reason to believe many common force fields are not ergodic. Systems with multiple-welled potentials are often non-ergodic since they only explore one of the relevant wells, and all torsion potentials are usually multi-welled (corresponding to rotating a bond around itself, which almost never happens). Systems with a large number of independent or mostly independent oscillations are rarely ergodic, since each individual oscillator has additional conserved quantities associated with it beyond the

overall system energy [18]. Many force fields to a first-order approximation assume bonds are harmonic, so all the bonds in a protein form a system of mostly independent harmonic oscillators that we do not expect to be ergodic. Thus we have good reason to suspect that force fields do not satisfy the ergodic hypothesis.

We will instead assume complete integrability, often considered the exact opposite of ergodicity. We will justify this assumption later. Here we motivate the conditions of complete integrability.

If we knew ahead of time that the motion of our system were periodic with respect to any one particular coordinate $x$, it would be trivial to show that spatial averages with respect to $x$ and time averages converge to each other; just integrate over a period. However, most MD simulations are not periodic with respect to any spatial coordinate, and this method does not allow you to average over any coordinate aside from $x$.

But consider the 1D harmonic oscillator

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

(setting constants equal to 1 for simplicity). It is well-known that the trajectory of a particle in this Hamiltonian is $(p(t), q(t)) = (r\cos\theta, r\sin\theta)$ for some fixed $r = \sqrt{2H}$ and linear function $\theta$. We can change coordinates from $(p, q)$ to $(r, \theta)$; now, $r$ is fixed over our trajectory and we can integrate with respect to $\theta$, a periodic coordinate. Here, the time average of a function $A$ is $\frac{1}{T}\int_0^T A(r\cos\theta(t), r\sin\theta(t))\,dt$ and the spatial average is $\frac{1}{2\pi}\int_0^{2\pi} A(r\cos\theta, r\sin\theta)\,d\theta$ which are clearly equal as long as $T$ is a multiple of the period. In this process, we have transformed our system from non-periodic coordinates to periodic coordinates, allowing us to compare spatial and time averages easily. We aim to repeat this process with more general MD simulations to create periodic coordinates.

The proof outline provided above for the 1-dimensional harmonic oscillator will be the motivation for the following chapter. We will find ways to construct periodic coordinates like $(r, \theta)$ and use them to establish rigorous bounds on sampling error for a subset of common MD force fields.

## 3.1 Integrable Systems and the Arnold-Liouville Theorem

In this section, we will define a completely integrable system, prove the Arnold-Liouville theorem, which characterizes completely integrable systems in terms of periodic coordinates as described previously, and discuss why complete integrability is a reasonable assumption to make about MD force fields.

### 3.1.1 Poisson Brackets and First Integrals

We first generalize the notion of a conserved quantity such as the energy of the harmonic oscillator.

**Definition 3.1.1.** A **first integral** of motion is a quantity conserved over the trajectory that is a function of only phase space coordinates (position and momentum) and is independent of time.

Energy, angular momentum, and linear momentum are common examples of first integrals. Quantities that are only conserved, i.e. independent of time, are generally called

integrals of motion. We'll need a slightly more useful characterization of first integrals. Consider functions $f, g : M \to \mathbb{R}$ on a symplectic manifold $M$. Then:

**Definition 3.1.2.** The **Poisson bracket** $\{f, g\}$ is a function $h : M \to \mathbb{R}$ such that $h = \omega(\overline{f}, \overline{g})$ where $\omega$ is the symplectic structure on $M$ and $\overline{f}, \overline{g}$ are the vector fields corresponding to $f, g$ [2].

**Example 3.1.3.** In our canonical basis, we may compute

$$\{f, g\} = \sum_{i=1}^{n} \frac{\partial g}{\partial p_i} \frac{\partial f}{\partial q_i} - \frac{\partial g}{\partial q_i} \frac{\partial f}{\partial p_i}.$$

The Poisson bracket is clearly linear and anti-symmetric by definition. The reason Poisson brackets are useful for computing first integrals is the following lemma:

**Lemma 3.1.4.** *A function f is a first integral iff we have* $\{f, H\} = 0$, *where H is the Hamiltonian.*

*Proof.* We have

$$\{f, H\} = \omega(\overline{f}, \overline{H}) = df(\overline{H})$$

by unwinding definitions and

$$\frac{d}{dt} f(\phi_t(x)) = df\left(\frac{d}{dt}\phi_t(x)\right) = df(\overline{H}(x))$$

by the Chain Rule. Thus $\{f, H\} = 0$ is equivalent to $\frac{d}{dt} f(\phi_t(x)) = 0$ always. □

We can also show that

$$\{f, g\} = \pounds_{\overline{g}} f$$

(where $\pounds$ is the Lie derivative) and the Jacobi identity holds for Poisson brackets, i.e.

$$\{f, \{g, h\}\} + \{g, \{h, f\}\} + \{h, \{f, g\}\} = 0.$$

Finally, the Poisson bracket is related to the Lie bracket by the identity

$$\{f, g\} = [\overline{f}, \overline{g}].$$

As a result, if $\{f, g\} = 0$, the flows with respect to $f$ and $g$ commute. We say that such $f$ and $g$ are **in involution**. These results are proven in Arnold [2].

### 3.1.2 Liouville's Theorem

Liouville's theorem gives us a very nice description of a manifold with multiple first integrals. The proof of this theorem will be the focus of this section.

**Theorem 3.1.5** (Liouville's Theorem). *Given a symplectic 2n-dimensional manifold M, suppose that we have n functions $f_1, \ldots, f_n$ such that $f_1 = H$, our Hamiltonian, $\{f_i, f_j\} = 0$ for all $i, j$ (in particular $f_i$ is a first integral for $i \neq 1$), and given a level set $M_{\vec{g}} = \{x \mid f_i(x) = g_i \forall i\}$ for some constants $g_1, \ldots, g_n$, the n 1-forms $df_1$ are linearly independent on $M_{\vec{g}}$, then we have the following:*

1. $M_{\vec{g}}$ is a smooth manifold invariant under phase flow by the Hamiltonian.

2. If $M_{\vec{g}}$ is compact and connected, it is diffeomorphic to the *n*-dimensional torus $\mathbb{T}^n$ with coordinates $\varphi_1, \ldots, \varphi_n \pmod{2\pi}$.

3. In these coordinates, the phase flow is **conditionally periodic**

$$\frac{d\vec{\phi}}{dt} = \vec{\omega}$$

for $\vec{\omega}$ a function only of $\vec{g}$.

We use the term conditionally periodic since the flow is only periodic if all the $\omega_i$s are related to each other by rational ratios. Otherwise, the flow is **quasiperiodic** [26]. The $M_{\vec{g}}$ are sometimes called **invariant tori**.

**Definition 3.1.6.** A system satisfying the conditions of Liouville's theorem is known as **completely integrable**.

*Proof of Liouville's Theorem from Arnold [2].* We start by proving the first statement. Since the *n* 1-forms $df_i$ are linearly independent, by the implicit function theorem, $M_{\vec{g}}$ must be a smooth *n*-dimensional submanifold.

**Lemma 3.1.7.** *The n vector fields $\overline{f}_i$ defined using the symplectic structure $\omega$ are tangent to $M_{\vec{g}}$, commute with each other, and are linearly independent.*

*Proof from Arnold [2].* Since we have an isomorphism from 1-forms to vector fields, linear independence of the $\overline{f}_i$ is immediately implied by linear independence of the $df_i$. The vector fields commuting follows from $\{f_i, f_j\} = 0$. Similarly, $\pounds_{\overline{f}_j} f_i = 0$, implying that the $\overline{f}_j$ are tangent to $M_{\vec{g}}$, as desired. $\square$

By Lemma 3.1.7, the $\overline{f}_i$ form a basis for $T_x M_{\vec{g}}$ at any $x \in M_{\vec{g}}$, implying that $\omega = 0$ on $T_x M_{\vec{g}}$ since it is 0 on all the $\overline{f}_i$. Further, we can define flows $\phi_{i,t}$ corresponding to the vector field $\overline{f}_i$ and a time $t$. These flows also commute and $M_{\vec{g}}$ is invariant under their flow. This proves the first statement of Liouville's theorem.

Now we prove the second statement. Assume $M_{\vec{g}}$ is compact and connected. Define an action of $\mathbb{R}^n$ on $M_{\vec{g}}$ by sending

$$\vec{t} \mapsto \phi_{\vec{t}} : M \to M, \phi_{\vec{t}} = \phi_{1,t_1} \phi_{2,t_2} \ldots \phi_{n,t_n}.$$

Commutativity of our flows implies that this is a group action. Given any point $x_0 \in M_{\vec{g}}$, we have a map $\phi : \mathbb{R}^n \to M_{\vec{g}}, \phi(\vec{t}) = \phi_{\vec{t}} x_0$.

By the implicit function theorem applied to the $\phi_{i,t}$, some neighborhood $V \subset \mathbb{R}^n$ of the origin maps to an open set $U$ containing $x_0$ diffeomorphically. We claim that $\phi$ is surjective; consider a curve from $x_0$ to any $x \in M$. We can cover this curve by a finite number of copies of $U$; each $U$ corresponds under $\phi$ to some $V$ and thus each point in the

curve corresponds to some shift in $V$. We now need to add up all of these shifts to get a $\vec{t}$ such that $\phi(\vec{t}) = x$.

However, $\phi$ is not injective since $M_{\vec{g}}$ is compact. Thus consider the stabilizer of $x_0$; this is a subgroup $\Gamma$. Further, since $\phi : V \to U$ is a diffeomorphism, $V \nsubseteq \Gamma$, implying that $\Gamma$ is discrete.

It is well-known that all discrete subgroups of $\mathbb{R}^n$ correspond to $\{ \sum_i m_i \vec{e}_i \mid m_i \in \mathbb{Z} \}$ for some set of $k$ linearly independent vectors $\vec{e}_i$ with $0 < k \leqslant n$. We use this to construct a diffeomorphism between $M_{\vec{g}}$ and $\mathbb{T}^k \times \mathbb{R}^{n-k}$. See the following diagram:

$$
\begin{array}{ccc}
\mathbb{R}^n & \xrightarrow{\ A\ } & \mathbb{R}^n \\
\downarrow{\scriptstyle p} & & \downarrow{\scriptstyle \phi} \\
\mathbb{T}^k \times \mathbb{R}^{n-k} & \xrightarrow{\ \widetilde{A}\ } & M_{\vec{g}}
\end{array}
$$

Take coordinates on $\mathbb{T}^k \times \mathbb{R}^{n-k}$ given by $(\varphi_1, \dots, \varphi_k, y_{k+1}, \dots, y_n)$. The $\varphi_i$ coordinates are taken mod $2\pi$. Define $p : \mathbb{R}^n \to \mathbb{T}^k \times \mathbb{R}^{n-k}$ to be the natural universal covering map that sends the first $k$ coordinates to their value mod $2\pi$. Note that by the above arguments $\phi : \mathbb{R}^n \to M_{\vec{g}}$ is also a universal covering map. Now define the points $\vec{h}_i \in \mathbb{R}^n$ with coordinates $\varphi_i = 2\pi, \varphi_j = 0$ for all $j \neq i$ and the isomorphism $A : \mathbb{R}^n \to \mathbb{R}^n, \vec{h}_i \mapsto \vec{e}_i$ (extend by linearity if necessary). This isomorphism descends to an isomorphism $\widetilde{A} : \mathbb{T}^k \times \mathbb{R}^{n-k} \to M_{\vec{g}}$ by the universal cover maps; since the universal cover maps are locally diffeomorphic, $\widetilde{A}$ is a diffeomorphism.

Recall that $M_{\vec{g}}$ is compact. $\mathbb{T}^k \times \mathbb{R}^{n-k}$ is compact iff $n = k$, implying that we actually have a diffeomorphism $M_{\vec{g}} \simeq \mathbb{T}^n$ and have constructed $n$ angular coordinates $\varphi_1, \dots, \varphi_n$. This proves the second statement of Liouville's theorem. The third statement of Liouville's theorem is also easy to see; note that by definition $\vec{\varphi} = A^{-1}\vec{t}$ and flowing along the Hamiltonian corresponds to $\frac{dt_1}{dt} = 1, \frac{dt_i}{dt} = 0$ by definition, so $\frac{d\varphi_i}{dt}$ must be some constant depending only on $\vec{g}$ as desired. $\qquad\square$

Liouville's theorem provides us with angular coordinates given a large number of first integrals, just as we saw in the motivating example of the harmonic oscillator.

### 3.1.3 Action-Angle Coordinates

We specialize to the case $M = \mathbb{R}^{2n}$ for this part.

Given a completely integrable system, we may by Liouville's theorem switch to coordinates $(\vec{f}, \vec{\varphi})$ where the flow corresponds to

$$
\frac{d\vec{f}}{dt} = 0, \frac{d\vec{\varphi}}{dt} = \vec{\omega}(\vec{f}).
$$

These new coordinates are not quite symplectic however, in the sense that we do not have $\omega = \sum df_i \wedge d\varphi_i$, which will prove very useful in bounding sampling error. We thus derive **action-angle coordinates** $(\vec{I}, \vec{\varphi})$, for which

$$
\omega = \sum_i dI_i \wedge d\varphi_i
$$

does hold. $\vec{I}$ is a function of $\vec{f}$ and thus $I_i$ is also a first integral of our Hamiltonian $H$ [2].

**Example 3.1.8.** For our harmonic oscillator, action-angle coordinates would be $I = H = \frac{p^2+q^2}{2}$. We then observe that

$$\mathrm{d}p \wedge \mathrm{d}q = r\,\mathrm{d}r \wedge \mathrm{d}\varphi = \mathrm{d}\left(\frac{r^2}{2}\right) \wedge \mathrm{d}\varphi$$

so these are indeed action-angle coordinates.

**Theorem 3.1.9** (Arnold-Liouville Theorem). *For our manifold $M = \mathbb{R}^{2n}$, there exist action-angle coordinates.*

*Proof from Arnold [2].* Let $\gamma_1, \ldots, \gamma_n$ be 1-dimensional cycles on $M_{\vec{g}}$ that form a basis of cycles. Define

$$I_i(\vec{g}) = \frac{1}{2\pi} \oint_{\gamma_i} \vec{p}\,d\vec{q}.$$

This integral does not depend on choice of $\gamma_i$ since given any two $\gamma, \gamma'$ we may define a cycle $\sigma$ such that $\partial\sigma = \gamma - \gamma'$ and then by Stokes's theorem

$$\oint_{\gamma} \vec{p}\,d\vec{q} - \oint_{\gamma'} \vec{p}\,d\vec{q} = \int\int_{\sigma} \sum_i \mathrm{d}p_i \wedge \mathrm{d}q_i = 0$$

since $\omega = 0$ on $M_{\vec{g}}$. The $I_i(\vec{g})$ are our action variables.

We now need to show that

$$\sum \mathrm{d}p_i \wedge \mathrm{d}q_i = \sum \mathrm{d}I_i \wedge \mathrm{d}\varphi_i.$$

Note that, as above,

$$S(x) = \int_{x_0}^{x} \vec{p}\,d\vec{q}$$

when restricted to $M_{\vec{g}}$ does not change under deformations of the path of integration. By the fundamental group, it can only change by values corresponding to $2\pi I_i$. Let $x_0 \in M_{\vec{g}}$ such that we can find a neighborhood where the $\vec{q}$ are coordinates on $M_{\vec{g}}$. In this neighborhood, we can define a function

$$S(\vec{I}, \vec{q}) = \int_{q_0}^{q} \vec{p}(\vec{I}, \vec{q})\,d\vec{q}.$$

It's immediate that

$$p_i = \frac{\partial S}{\partial q_i}.$$

Define

$$\frac{\partial S}{\partial I_i} = \varphi_i.$$

Thus, at least locally, we have our desired equality by taking second derivatives of $S$. This local equality can be extended globally by observing that the discrepancy in defining $S(\vec{I}, \vec{q})$ corresponds exactly to the discrepancy in precisely defining $\vec{\varphi}$.     $\square$

### 3.1.4 Integrability and MD Force Fields

We should next consider whether integrability applies to a significant number of force fields. As described previously, the 1-dimensional harmonic oscillator, and in fact any system built by adding independent higher-dimensional harmonic oscillators, is completely integrable [26]. Unfortunately, just as with ergodicity, it is extremely difficult to prove that any force field generates a completely integrable Hamiltonian. Further, just as with ergodicity, the set of integrable Hamiltonians is also meager in the set of all Hamiltonians [36].

So why do we consider integrable systems as a useful approximation to the behavior of MD simulations, in contrast with ergodicity? The key mathematical result is the following:

**Theorem 3.1.10** (Kolmogorov-Arnold-Moser (KAM) Theorem)**.** *Given an unperturbed non-degenerate (i.e.* $\det \left| \frac{\partial \omega}{\partial I} \right| \neq 0$*) completely integrable analytic Hamiltonian* $H_0$*, adding small conservative Hamiltonian perturbations* $H'$ *only slightly deforms non-resonant invariant tori* $M_{\vec{g}}$*. In the phase space of the perturbed system, we also have invariant tori on which the trajectories are conditionally-periodic. These invariant tori form a majority of phase space of the perturbed system, i.e. the measure of the complement of their union is small when their perturbation is small.*

The proof of the KAM theorem can be found in Arnold [2]. The KAM theorem tells us that the fundamental structural fact about completely integrable systems, the existence of invariant tori and conditionally periodic trajectories, is largely preserved upon small perturbations. Thus, while small perturbations of completely integrable systems may result in non-integrable systems, we still have our invariant tori and can still draw conclusions about the behavior of our system in the majority of phase space. This is contrast to ergodicity, where we have no reason to believe the structure of ergodic systems is maintained even under small perturbations from ergodic Hamiltonians.

There are a couple of other reasons to believe that MD force fields are completely integrable or close to completely integrable. Intuitively speaking, we can think of the result of Liouville's theorem as claiming that there are a number of distinct frequencies under which our system moves, and this does correspond closely to how biomolecular simulations behave; it's easy to draw a distinction between the frequencies associated with bond stretches, bonds bending, torsion, and non-bonded forces. Thus, there is reason to believe that working with completely integrable systems will give us a good understanding of convergence properties of molecular dynamics simulations [9].

Finally, Cances et al. [9] have previously found that the convergence properties we prove in subsequent sections hold up empirically for real MD simulations. This is weak empirical justification for the fact that force fields are in fact completely integrable or close to completely integrable. We will present these empirical results in Section 3.2.4.

Despite these arguments, it would obviously be better to prove the convergence of time and spatial averages and bound sampling error independent of any assumptions on the force field. Unfortunately, this remains an open problem, so we focus here on what has already been proven [31, 7, 33].

## 3.2 Sampling Error

Suppose we're given an observable $A(\vec{p}, \vec{q})$. Define the spatial average

$$\langle A \rangle = \frac{\int_M A(\vec{p}, \vec{q}) \, d\mu(\vec{p}, \vec{q})}{\int_M d\mu(\vec{p}, \vec{q})} \tag{3.1}$$

where $M$ is the manifold corresponding to accessible phase space (conserving energy and whatever other first integrals we have) and $d\mu$ is the canonical measure inherited from all of phase space $\mathbb{R}^{2n}$. Define the time average

$$\langle A \rangle (T) = \frac{1}{T} \int_0^T A(\vec{p}(t), \vec{q}(t)) \, dt \tag{3.2}$$

given a trajectory $\vec{p}(t), \vec{q}(t)$.

In this section, we will show that $\lim_{T \to \infty} \langle A \rangle (T) = \langle A \rangle$. We will also investigate sampling error, the error associated with bounding the left hand side at some finite $T$. As explained previously, we will work on invariant tori, associated with completely integrable Hamiltonians and most of the phase space of Hamiltonians slightly perturbed from completely integrable ones.

### 3.2.1 Equivalence of Spatial and Time Averages

We will need one additional condition on our system to ensure that spatial and time averages are equal. Call the elements of $\vec{\omega}$ our frequencies, and let them be **independent** if they are linearly independent over the field of rational numbers, i.e. if $\vec{k} \in \mathbb{Q}^n$ and $\vec{k} \cdot \vec{\omega} = 0$, then $\vec{k} = 0$. This is a reasonable condition to place for realistic systems, since it's highly unlikely that two distinct frequencies of a natural system would be related by a rational number (slight perturbations would make this untrue). One can extend these results to show that given $r$ linear dependences, our invariant torus reduces to an invariant torus of dimension $n - r$ with $n - r$ independent frequencies and conditionally periodic motion, at which point our results still apply as usual [2]. Systems of identical particles can thus be treated by reducing the number of independent frequencies.

**Theorem 3.2.1.** *If our Hamiltonian is completely integrable and our frequencies are independent, then*

$$\lim_{T \to \infty} \langle A \rangle (T) = \langle A \rangle$$

*for any continuous observable A.*

Our primary strategy for proving this result and subsequent similar results will be the Fourier expansion. We will briefly review the Fourier expansion for a real continuous function $f$ over a multidimensional torus $\mathbb{T}^n$. We have

$$f = \sum_{\vec{k} \in \mathbb{Z}^n} \hat{f}(\vec{k}) e^{i\vec{k} \cdot \vec{\varphi}}$$

where

$$\hat{f}(\vec{k}) = \frac{1}{(2\pi)^n} \int_{\mathbb{T}^n} f(\vec{\varphi}) e^{-i\vec{k}\cdot\vec{\varphi}} \, d\vec{\varphi}.$$

Convergence can be proven by noting that the exponentials form an orthonormal basis for a Hilbert space or by applying the Stone-Weierstrass theorem [2].

*Proof of Theorem 3.2.1 modified from Arnold [2].* Work on our invariant torus with coordinates $\vec{\varphi}$. Then the spatial average from Equation 3.1 reduces to the integral

$$\langle A \rangle = \frac{1}{(2\pi)^n} \int_0^{2\pi} \dots \int_0^{2\pi} f(\vec{\varphi}) \, d\varphi_1 \dots d\varphi_n.$$

The time average from Equation 3.2 is now

$$\langle A \rangle (T) = \frac{1}{T} \int_0^T A(\vec{\varphi}(t)) \, dt = \frac{1}{T} \int_0^T A(\vec{\varphi}_0 + \vec{\omega}t) \, dt.$$

We'll first prove the result for exponentials $A = e^{i\vec{k}\cdot\vec{\varphi}}$. In this case, if $\vec{k} = 0$, then $A = 1$ so $\langle A \rangle (T) = \langle A \rangle = 1$ always. If not, then our spatial average is an integral over an entire period of a complex exponential, so $\langle A \rangle = 0$. Our time average is

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T e^{i(\vec{k}\cdot(\vec{\varphi}_0 + \vec{\omega}t))} \, dt = \lim_{T\to\infty} \frac{e^{i\vec{k}\cdot\vec{\varphi}_0}}{i\vec{k}\cdot\vec{\omega}} \frac{e^{iT(\vec{k}\cdot\vec{\omega})} - 1}{T} = 0$$

as desired.

Our proof above immediately implies the result for trigonometric polynomials, since both the time and spatial averages are linear.

Now consider an arbitrary observable $A$ and take its Fourier expansion. For fixed $\epsilon > 0$, by convergence of the Fourier series, we may pick a partial sum (a trigonometric polynomial) $P$ such that $|A - P| < \frac{\epsilon}{3}$. Therefore

$$|\langle A \rangle - \langle P \rangle| = \left| \frac{1}{(2\pi)^n} \int_{\mathbb{T}^n} (A - P) \, d\vec{\varphi} \right| \leqslant \frac{1}{(2\pi)^n} \int_{\mathbb{T}^n} |A - P| \, d\vec{\varphi} \leqslant \frac{\epsilon}{3}$$

and similarly $|\langle A \rangle (T) - \langle P \rangle (T)| \leqslant \frac{\epsilon}{3}$. Pick $T$ large enough so by the above argument $|\langle P \rangle - \langle P \rangle (T)| < \frac{\epsilon}{3}$. Combining these, we get $|\langle A \rangle - \langle A \rangle (T)| < \epsilon$. By the definition of limit, we now have $\lim_{T\to\infty} \langle A \rangle (T) = \langle A \rangle$ as desired. □

### 3.2.2 Bounding Sampling Error

We'd now like to modify the above proof to determine the rate at which sampling error converges to 0. To do so, we will require one additional condition on the potential values of $\vec{k} \cdot \vec{\omega}$.

We assume **Siegel's diophantine condition**, i.e. that there exist $\gamma, \nu > 0$ such that for all $\vec{k} \in \mathbb{Z}^n$ nonzero we have

$$|\vec{k} \cdot \vec{\omega}| > \gamma |\vec{k}|^{-\nu}. \tag{3.3}$$

This condition intuitively just means that our frequencies are not particularly close to being linearly independent [9]. It can be shown that the set of frequencies not satisfying this condition for $\nu > n - 1$ has Lebesgue measure $\mathcal{O}(\gamma)$ which tends to 0 as $\gamma \to 0$. Thus almost all frequencies satisfy this condition for some $\gamma$ [26]. We will also assume our observable is real analytic (reasonable for any physically realistic observable).

**Theorem 3.2.2.** *Under the same conditions as before and Siegel's diophantine condition, there exists a constant c such that*

$$|\langle A \rangle (T) - \langle A \rangle| \leqslant \frac{c}{T}$$

*for any analytic observable A.*

*Proof modified from Cances et al. [9].* Take the Fourier expansion of $A$ as

$$A(\vec{\varphi}) = \sum_{\vec{k} \in \mathbb{Z}^n} \hat{A}(\vec{k}) e^{i\vec{k} \cdot \vec{\varphi}}.$$

Note that by an extension of the Paley-Wiener theorem, analyticity of $A$ implies that the Fourier coefficients exponentially decay, i.e. we have $\hat{A}(\vec{k}) \leqslant C e^{-D|\vec{k}|}$ for some positive constants $C, D$ independent of $T$.

To compute the spatial and time averages, we will want to switch integrals and infinite sums of the Fourier expansion freely in the subsequent steps. We may justify this by using the dominated convergence theorem; note that

$$|\hat{A}(\vec{k}) e^{i\vec{k} \cdot \vec{\varphi}}| \leqslant \hat{A}(\vec{k}) \leqslant C e^{-D|\vec{k}|}$$

by the above, so our sum is dominated by a convergent sum and we may apply the dominated convergence theorem to freely bring integrals inside the Fourier expansion.

As a result, we have

$$\langle A \rangle = \sum_{\vec{k} \in \mathbb{Z}^n} \left\langle \hat{A}(\vec{k}) e^{i\vec{k} \cdot \vec{\varphi}} \right\rangle = \hat{A}(\vec{0})$$

by the same reasoning as in the proof of Theorem 3.2.1. Further,

$$\langle A \rangle (T) = \sum_{\vec{k} \in \mathbb{Z}^n} \left\langle \hat{A}(\vec{k}) e^{i\vec{k} \cdot \vec{\varphi}} \right\rangle = \hat{A}(\vec{0}) + \frac{1}{T} \sum_{\vec{k} \in \mathbb{Z}^n, \vec{k} \neq \vec{0}} \hat{A}(\vec{k}) \frac{e^{i\vec{k} \cdot \vec{\varphi}_0} \left( e^{iT(\vec{k} \cdot \vec{\omega})} - 1 \right)}{i\vec{k} \cdot \vec{\omega}}.$$

Using Siegel's diophantine condition (Equation 3.3), we may now bound the error by

$$|\langle A \rangle (T) - \langle A \rangle| \leqslant \frac{1}{T} \sum_{\vec{k} \in \mathbb{Z}^n, \vec{k} \neq \vec{0}} \left| \hat{A}(\vec{k}) \frac{e^{i\vec{k} \cdot \vec{\varphi}_0} \left( e^{iT(\vec{k} \cdot \vec{\omega})} - 1 \right)}{i\vec{k} \cdot \vec{\omega}} \right| \leqslant \frac{1}{T} \sum_{\vec{k} \in \mathbb{Z}^n, \vec{k} \neq \vec{0}} \frac{2C}{\gamma} |\vec{k}|^\nu e^{-D|\vec{k}|}.$$

By the integral test, this sum is convergent, thus we may take $c = \sum_{\vec{k} \in \mathbb{Z}^n, \vec{k} \neq \vec{0}} \frac{2C}{\gamma} |\vec{k}|^\nu e^{-D|\vec{k}|}$ and we have the desired. $\qquad \square$

### 3.2.3 Reducing Sampling Error with Filter Functions

It's reasonable to ask whether it is possible to improve this bound by slightly modifying the computation of time averages. We obviously do not want to increase computational time any further, so the only real possible modification is by adding a positive **filter function** $\xi(t) : [0,1] \to \mathbb{R}^{\geq 0}$ and computing the average

$$\langle A \rangle_\xi (T) = \frac{\int_0^T A(\varphi(t))\xi\left(\frac{t}{T}\right) \, dt}{\int_0^T \xi\left(\frac{t}{T}\right) \, dt} \tag{3.4}$$

instead. Our previous case was effectively $\xi(t) = 1$. The intuitive idea behind the filter is to place unequal weights on the value of our function at certain points in the trajectory and compute a weighted average in an attempt to improve convergence. We will later show an example of a particularly nice filter function.

**Theorem 3.2.3.** *Under the same conditions as before, and given that $\xi$ is $(d+1)$st times differentiable with $\xi^{(j)}(0) = \xi^{(j)}(1) = 0$ for $j = 0, \ldots, d-1$, then there exist constants $c_0$ and $R$ not dependent on $d$ such that*

$$|\langle A \rangle_\xi (T) - \langle A \rangle| \leqslant \frac{1}{T^{d+1}} \left( c_0 R^{d+1} (\nu(d+1))! \frac{|\xi^{(d)}(0)| + |\xi^{(d)}(1)| + ||\xi^{(d+1)}||_{L^1}}{||\xi||_{L^1}} \right).$$

As usual, the $L^1$ norm is defined by $||\xi||_{L^1} = \int_0^1 \xi(t) \, dt$. We pick $\nu$ integer in Siegel's diophantine condition so the factorial makes sense.

*Proof from Cances et al. [9].* We have $\langle A \rangle = \hat{A}(\vec{0})$ and expanding Equation 3.4,

$$\langle A \rangle_\xi (T) = \hat{A}(\vec{0}) + \frac{1}{T||\xi||_{L^1}} \sum_{\vec{k} \in \mathbb{Z}^n, \vec{k} \neq \vec{0}} \hat{A}(\vec{k}) e^{i\vec{k} \cdot \vec{\varphi}_0} \int_0^T e^{it(\vec{k} \cdot \vec{\omega})} \xi\left(\frac{t}{T}\right) \, dt$$

by similar reasoning as in the proof of Theorem 3.2.2. We now want to estimate this integral.

Integrate by parts $d$ times. We obtain

$$\int_0^T e^{it(\vec{k} \cdot \vec{\omega})} \xi\left(\frac{t}{T}\right) \, dt = \frac{(-1)^d}{(Ti(\vec{k} \cdot \vec{\omega}))^d} \int_0^T \xi^{(d)}\left(\frac{t}{T}\right) e^{it(\vec{k} \cdot \vec{\omega})} \, dt$$

since the boundary terms all vanish by our given conditions. Integrating by parts one more time, we get

$$\int_0^T e^{it(\vec{k} \cdot \vec{\omega})} \xi\left(\frac{t}{T}\right) \, dt = \frac{(-1)^d}{(Ti(\vec{k} \cdot \vec{\omega}))^{k+1}} T \left( \xi^{(d)}\left(\frac{t}{T}\right) e^{it(\vec{k} \cdot \vec{\omega})} \Big|_0^T \right) - \frac{(-1)^d}{(Ti(\vec{k} \cdot \vec{\omega}))^{d+1}} \int_0^T \xi^{(d+1)}\left(\frac{t}{T}\right) e^{it(\vec{k} \cdot \vec{\omega})} \, dt$$

$$\leqslant \frac{1}{|\vec{k} \cdot \vec{\omega}|^{d+1} T^d} \left( |\xi^{(d)}(0)| + |\xi^{(d)}(1)| + ||\xi^{(d+1)}||_{L^1} \right)$$

We now can plug this back into our original expression. This yields

$$|\langle A \rangle_\xi (T) - \langle A \rangle| \leqslant \frac{1}{T^{d+1}} \frac{|\xi^{(d)}(0)| + |\xi^{(d)}(1)| + ||\xi^{(d+1)}||_{L^1}}{||\xi||_{L^1}} \sum_{\vec{k} \in \mathbb{Z}^n, \vec{k} \neq \vec{0}} \frac{|\hat{A}(\vec{k})|}{|\vec{k} \cdot \vec{\omega}|^{k+1}}$$

and it remains to bound this last sum. Applying Siegel's diophantine condition (Equation 3.3) and the Fourier expansion bound for analyticity, we have

$$\sum_{\vec{k} \in \mathbb{Z}^n, \vec{k} \neq \vec{0}} \frac{|\hat{A}(\vec{k})|}{|\vec{k} \cdot \vec{\omega}|^{d+1}} \leqslant \sum_{\vec{k} \in \mathbb{Z}^n, \vec{k}0} \frac{C}{\gamma^{d+1}} |\vec{k}|^{\nu(d+1)} e^{-D|\vec{k}|} \leqslant \sum_{\vec{k} \in \mathbb{Z}^n, \vec{k} \neq \vec{0}} \frac{C}{\gamma^{d+1} \left(\frac{D}{2}\right)^{\nu(d+1)}} (\nu(d+1))! e^{-\frac{D}{2}|\vec{k}|}$$

using the fact that $x^n \leqslant n! e^x$ which is true by Taylor series for positive $x$. This sum converges for the same reasons as before, so we may now pick $c_0 = C \sum_{\vec{k} \in \mathbb{Z}^n, \vec{k} \neq \vec{0}} e^{-\frac{D}{2}|\vec{k}|}$, $R = \frac{1}{\gamma(\frac{D}{2})^\nu}$ to get the desired. $\qquad\square$

A particularly useful filter function is

$$\xi(x) = e^{-\frac{1}{x(1-x)}}$$

which satisfies the condition above for all $d \in \mathbb{N}$. As shown in [9], there exist constants $\mu, \beta, \delta > 0$ such that

$$||\xi^{(d)}|| \leqslant \mu \beta^d d^{\delta d}, ||\xi^{(d)}||_{L^\infty} \leqslant \mu \beta^d d^{\delta d}$$

where the $L^\infty$ norm is the supremum over $[0, 1]$. Letting $c_0 = c\mu, r_1 = R\beta$, we now have

$$|\langle A \rangle_\xi (T) - \langle A \rangle| \leqslant c_1 \left(\frac{r_1}{T}\right)^{d+1} (d+1)^{\delta(d+1)} (\nu(d+1))! \leqslant c_1 \left(\frac{r_1 \nu^\nu}{T}\right)^{d+1} (d+1)^{(\delta+\nu)(d+1)}$$

for all $d$. The RHS is a function of $d$ which can be minimized; computing this, we get

$$|\langle A \rangle_\xi (T) - \langle A \rangle| \leqslant c_1 \exp\left(-\frac{\delta + \nu}{e} \left(\frac{T}{r_1 \nu^\nu}\right)^{\frac{1}{\delta+\nu}}\right).$$

Thus we get an exponentially decaying bound on the size of our error, which is a significant improvement over the naive results without a filter function [9].

The primary disadvantage of use of a filter function is that such methods have not been implemented in most standard MD packages, since most computational chemists think improvements on convergence of simulations are relatively unimportant. However, use of the filter function described above has led to significant improvements when tested [8]; we will now examine some numerical results that demonstrate our findings above.

### 3.2.4   Numerical Simulations

We now provide some empirical evidence to support the above results, in particular the claim that MD simulations are close to completely integrable. If they are in fact roughly completely integrable, we expect convergence rates of observables to be similar to those predicted by Theorem 3.2.3; it is still possible we will see good convergence for systems that are not completely integrable. We ignore time discretization error, since as shown in the next chapter it is relatively small for small time steps, and force-field error by assuming our force fields are completely accurate. We will look at cases where the system is essentially analytically solvable by other means.

The simulations presented in this section were carried out by Cances et al. [8]. They use $k$ where we have used $d$ above, so the figures are labeled accordingly.

We begin with a system of 5 harmonic oscillators

$$H = \sum_{i=1}^{5} \frac{p_i^2}{2} + \omega_i^2 \frac{q_i^2}{2}.$$

This system normally satisfies all the conditions in Theorem 3.2.3, but we make it violate Siegel's diophantine condition by picking frequencies $\omega_i$ that are rational numbers. The observable $A = p_4^2$ in this simulation. In Figure 3.2.1a, we see the predicted convergence bound; a least-squares fit verifies a decay of $\mathcal{O}(T^{-0.978})$ without any filter function and of $\mathcal{O}(T^{-3.87})$ with a filter function with $d = 3$, supporting our results [8].

The next potential is a Lennard-Jones potential that models van der Waals dispersion; this term would be very common in any MD simulation. We have $M = 288$ particles with

$$H = \sum_{i=1}^{M} \frac{\vec{p}_i^2}{2} + \sum_{i=1}^{M} \sum_{j>i} \left( \frac{4}{|\vec{q}_j - \vec{q}_i|^{12}} - \frac{4}{|\vec{q}_j - \vec{q}_i|^6} \right).$$

The initial conditions approximate a solid-phase system and the desired observable is the pressure $P$ (which can be written as a function of position and momentum). This system is not completely integrable since there are not enough first integrals, but we see in Figure 3.2.1b that it still satisfies the results of Theorem 3.2.3. Specifically, without any filter function we have a decay of $\mathcal{O}(T^{-0.996})$ and with a filter function with $d = 3$ we have a decay of $\mathcal{O}(T^{-3.76})$. The fact that an MD simulation satisfies our theorem without the completely integrable hypothesis is very promising for the applicability of our results in general [8].

The last potential under simulation was an alkane, or a chain of bonded atoms with potentials as outlined in Section 1.1.2.1. The observable is the end-to-end distance with a chain of $M = 40$ particles. This system is once again not completely integrable, but at low temperatures ($T = 2.13 \, \text{K}$) we find that our system exhibits the desired convergence properties in Figure 3.2.1c. In particular, without a filter function we have a decay of $\mathcal{O}(T^{-0.98})$ and with a filter function with $d = 3$ we have a decay of $\mathcal{O}(T^{-4.21})$ [8].
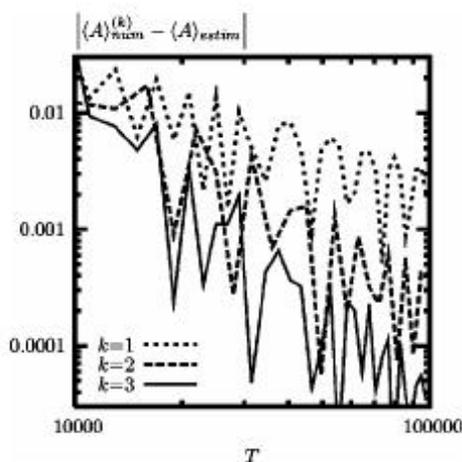
At high temperatures ($T = 135 \, \text{K}$) this no longer holds; our decay, with or without a filter function, is approximated as $\mathcal{O}(T^{-\frac{1}{2}})$. This result seems to hold in general for a large number of not completely integrable systems. One potential explanation for the failure
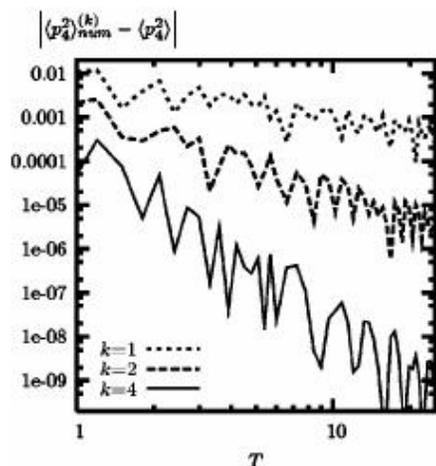
(a) Harmonic Oscillator
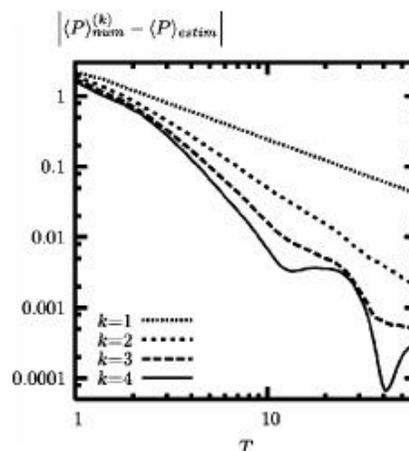
(b) Lennard-Jones

(c) Alkane, Low Temperature
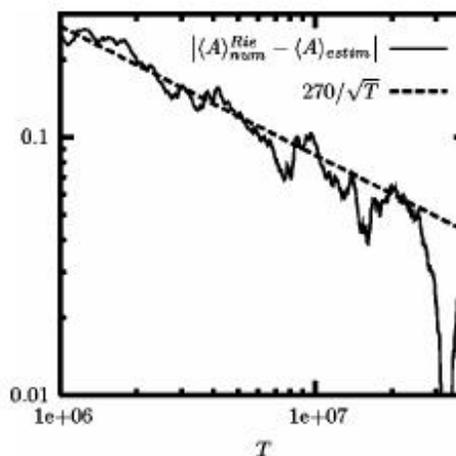
(d) Alkane, High Temperature

Figure 3.2.1: Bounds on Sampling Error in MD Simulation. Figures taken from Cances et al. [8]; $k$ refers to $d$, the expected bound on sampling error size. We see that Theorem 3.2.3 holds in some generality, as in (a) Siegel's diophantine condition fails and in (b) and (c) the system is not completely integrable, but we still have the desired convergence properties. However, in (d) we find that the Theorem no longer holds.

of our theorem is that our system now has many accessible energy wells, corresponding to different conformers of the alkane, and these are all accessible to the high temperature, which means the system has high kinetic/potential energy. Exploring these wells takes time, since the transition between wells takes time and is not energetically favorable. One potential way around this problem is by sampling a large number of conformations in different wells. While this is practical for the alkane, it is not for larger molecules that would also pose similar problems (since the number of wells grows exponentially). Proving this result mathematically and working around it for larger molecules remains an open question [8].

# Chapter 4

# Perturbation Theory and Time Discretization Error

Our next goal is to understand time discretization error, the error associated with using the Störmer-Verlet method instead of following the exact trajectory. Our primary tools for this analysis will be backwards error analysis and the Arnold-Liouville theorem on completely integrable systems, both of which we have already introduced. Backwards error analysis tells us that the trajectory from the Störmer-Verlet method is associated with a modified Hamiltonian that is a small perturbation from the original Hamiltonian of our system. Thus our goal for this chapter will be to understand perturbation theory on completely integrable systems and apply those results to the modified Hamiltonian of the Störmer-Verlet method.

## 4.1 Perturbation Theory for Integrable Systems

In this section, we will develop perturbation theory for integrable systems and understand how the invariant tori from the Arnold-Liouville theorem are almost preserved under small perturbations of the Hamiltonian.

Suppose we have a completely integrable Hamiltonian associated with our original force field $H_0(\vec{I})$ in action-angle coordinates $(\vec{I}, \vec{\varphi})$. We add a perturbation $H_1(\vec{I}, \vec{\varphi})$ such that our new Hamiltonian is

$$H(\vec{I}, \vec{\varphi}) = H_0(\vec{I}) + \epsilon H_1(\vec{I}, \vec{\varphi}) \tag{4.1}$$

for some small $\epsilon$. Assume that the perturbation $H_1$ is bounded (but may depend on $\epsilon$) and that all Hamiltonians are analytic.

We would like to change coordinates symplectically to a different coordinate system that looks more like action-angle coordinates for our new perturbative coordinate system. Specifically, we want to change coordinate $(\vec{I}, \vec{\varphi}) \mapsto (\vec{J}, \vec{\theta})$ such that the symplectic structure of our manifold is preserved, and the Hamiltonian depends on only $\vec{J}$ up to order $\mathcal{O}(\epsilon^N)$ for some $\epsilon$.

### 4.1.1 Linstedt-Poincaré Series

In order to do this, just as in the proof of Theorem , we use a generating function

$$S(\vec{J}, \vec{\varphi}) = \vec{J} \cdot \vec{\varphi} + \epsilon S_1(\vec{J}, \vec{\varphi}) + \epsilon^2 S_2(\vec{J}, \vec{\varphi}) + \ldots + \epsilon^{N-1} S_{N-1}(\vec{J}, \vec{\varphi}). \tag{4.2}$$

and then require $I_i = \frac{\partial S}{\partial \varphi_i}, \theta_i = \frac{\partial S}{\partial J_i}$ to enforce symplecticity. This expansion is known as a **Linstedt-Poincaré series**.

**Example 4.1.1.** Let's explicitly compute $N = 2$ [26]. Our generating function will look like

$$S(\vec{J}, \vec{\varphi}) = \vec{J} \cdot \vec{\varphi} + \epsilon S_1(\vec{J}, \vec{\varphi})$$

for some $S_1$. Let the new Hamiltonian be

$$K(\vec{J}, \vec{\theta}) = H(\vec{I}, \vec{\varphi}) = H_0(\vec{J}) + \epsilon K_1(\vec{J}, \vec{\theta}) + \dots$$

and we require that $K_1$ depend only on $\vec{J}$. Using Equation 4.1, we have

$$H(\vec{I}, \vec{\varphi}) = H\left(\vec{J} + \epsilon \frac{\partial S_1}{\partial \vec{\varphi}}(\vec{J}, \vec{\varphi}), \vec{\varphi}\right) = H_0(\vec{J}) + \epsilon \left(\frac{\partial H_0}{\partial \vec{I}}(\vec{J}) \cdot \frac{\partial S_1}{\partial \vec{\varphi}}(\vec{J}, \vec{\varphi}) + H_1(\vec{J}, \vec{\varphi})\right) + \dots. \tag{4.3}$$

Now define $\omega_0(\vec{J}) = \frac{\partial H_0}{\partial \vec{I}}(\vec{J})$ for convenience. We know that the quantity in parentheses in Equation 4.3 must be independent of $\vec{\varphi}$ so it must be equivalent to its average over $\vec{\varphi}$. This gives us the equation

$$\omega_0(\vec{J}) \cdot \frac{\partial S_1}{\partial \vec{\varphi}}(\vec{J}, \vec{\varphi}) + H_1(\vec{J}, \vec{\varphi}) = \frac{1}{(2\pi)^n} \int_{\mathbb{T}^n} H_1(\vec{J}, \vec{\varphi}) \, d\vec{\varphi}. \tag{4.4}$$

Apply a Fourier decomposition at each fixed $\vec{J}$ (so on a torus)

$$S_1(\vec{J}, \vec{\varphi}) = \sum_{\vec{k} \in \mathbb{Z}^n} \hat{S}_1(\vec{J}, \vec{k}) e^{i\vec{k} \cdot \vec{\varphi}}, \ H_1(\vec{J}, \vec{\varphi}) = \sum_{\vec{k} \in \mathbb{Z}^n} \hat{H}_1(\vec{J}, \vec{k}) e^{i\vec{k} \cdot \vec{\varphi}}.$$

The only surviving term on the RHS of Equation 4.4 is the $\vec{0}$th term. Thus, we get for $\vec{k} \neq 0$ that

$$\hat{S}_1(\vec{J}, \vec{k}) = -\frac{\hat{H}_1(\vec{J}, \vec{k})}{i\omega_0(\vec{J}) \cdot \vec{k}}.$$

Assume Siegel's diophantine condition (Equation 3.3) as before. Then if $H_1$ is analytic, the remainder term is bounded and this Fourier expansion converges. We can approximate the perturbation $H_1$ up to $\mathcal{O}(\epsilon^2)$ by a trigonometric polynomial of degree $m \sim |\log \epsilon|$ if we choose and then work entirely with trigonometric polynomials.

The general case is computed similarly [26, 41]. Taylor expanding around $\vec{I}$ for convenience, we get equations of the form

$$\omega_0(\vec{J}) \cdot \frac{\partial S_j}{\partial \vec{\varphi}}(\vec{J}, \vec{\varphi}) + K_j(\vec{J}, \vec{\varphi}) = \frac{1}{(2\pi)^n} \int_{\mathbb{T}^n} K_j(\vec{J}, \vec{\varphi}) \, d\vec{\varphi} \tag{4.5}$$

for $j = 1, \dots, N - 1$. Here, $K_1 = H_1$ as above,

$$K_2 = \frac{1}{2} \frac{\partial^2 H_0}{\partial \vec{I}^2} \left(\frac{\partial S_1}{\partial \vec{\varphi}}, \frac{\partial S_1}{\partial \vec{\varphi}}\right) + \frac{\partial H_1}{\partial \vec{I}} \cdot \frac{\partial S_1}{\partial \vec{\varphi}}$$

and in general $K_j$ is a sum of terms of the form

$$\frac{1}{i!} \frac{\partial^i H_{k_0}}{\partial \vec{I}^i} \left(\frac{\partial S_{k_1}}{\partial \vec{\varphi}}, \dots, \frac{\partial S_{k_i}}{\partial \vec{\varphi}}\right) \tag{4.6}$$

where $k_0 + \ldots + k_i = j$ and $k_0 = 0, 1$. This can be seen by an explicit Taylor expansion; multiple derivatives are interpreted as multilinear maps that take the given values as arguments.

These equations can be solved using Fourier expansions explicitly. Each Fourier expansion will converge as before assuming analyticity of $H_1$ and Siegel's diophantine condition. However, the entire series looks like

$$K(\vec{J}, \vec{\theta}) = H_0(\vec{J}) + \epsilon K_1(\vec{J}) + \epsilon^2 K_2(\vec{J}) + \ldots + \epsilon^{N-1} K_{N-1}(\vec{J}) + \epsilon^N R_N(\vec{J}, \vec{\varphi}) \tag{4.7}$$

with some remainder term $R_N$. Convergence of this series as $N \to \infty$ is not obvious; rather than dealing with these convergence issues, we will in general truncate $H_1$ to some trigonometric polynomial and thus deal with trigonometric polynomials $K_j$, $S_j$ [26].

## 4.1.2 Bounding the Generating Function

We'd first like to see how close the trajectory remains to conditionally periodic motion under a small perturbation. To do this, we'll need to begin with a few technical lemmas. The bounds that follow are obviously not very sharp, but they are sufficient for proving our final result.

Fix $\rho > 0$. Complexify the torus $\mathbb{T}^n$ to $U_\rho$ by defining

$$U_\rho := \{\vec{\varphi} \in \mathbb{T}^n + i\mathbb{R}^n \mid ||\operatorname{Im}\vec{\varphi}|| < \rho\}.$$

Here $|| \cdot ||$ refers to the max norm. We define a functional norm for a bounded analytic function $f$ on $U_\rho$ by

$$||f||_\rho = \sup_{\vec{\varphi} \in U_\rho} |F(\varphi)|, \left\|\frac{\partial f}{\partial \vec{\varphi}}\right\|_\rho = \sum_{j=1}^{n} \left\|\frac{\partial f}{\partial \varphi_j}\right\|_\rho.$$

**Lemma 4.1.2.** *Suppose $\vec{\omega} \in \mathbb{R}^n$ satisfies Siegel's diophantine condition (Equation 3.3). Let g be a bounded real-analytic function on $U_\rho$ and let $\overline{g} = \frac{1}{(2\pi)^n} \int_{\mathbb{T}^n} g \, d\vec{\varphi}$ be the spatial average over $\mathbb{T}^n$. Then the equation*

$$\omega \cdot \frac{\partial f}{\partial \vec{\varphi}} + g = \overline{g}$$

*has a unique real-analytic bounded solution f on $U_\rho$ with zero average $\overline{f} = 0$.*

*Fix positive $\delta < \min(\rho, 1)$. Then f is bounded on $U_{\rho-\delta}$ by*

$$||f||_{\rho-\delta} \leqslant \kappa_0 \delta^{-\alpha+1} ||g||_\rho, \left\|\frac{\partial f}{\partial \vec{\varphi}}\right\|_{\rho-\delta} \leqslant \kappa_1 \delta^{-\alpha} ||g||_\rho$$

*where $\alpha = \nu + n + 1, \kappa_0 = \gamma^{-1} 8^n 2^\nu \nu!, \kappa_1 = \gamma^{-1} 8^n 2^{\nu+1} (\nu+1)!$.*

The proof of this lemma is by Fourier expansion of the differential equation; it may be found in Hairer et al. [26]. We'll use this lemma to get bounds on the generating function $S_j$ and the remainder terms.

**Lemma 4.1.3.** *Let $H_0, H_1$ be real-analytic and bounded by $M$ on $B(\vec{J}^*, r)$, the ball of radius $r$ around $\vec{J}^* \in \mathbb{R}^n$, and on $B(\vec{J}^*, r) \times U_\rho$, respectively. Suppose $\omega(\vec{J}^*) = \frac{\partial H_0}{\partial \vec{I}}(\vec{J}^*)$ satisfies Siegel's diophantine condition (Equation 3.3). Then the coefficients of the*

$$\left\| \frac{\partial S_j}{\partial \vec{\varphi}}(\vec{J}^*, \cdot) \right\|_{\frac{\rho}{2}} \leqslant C_0 (C_1 j^\alpha)^{j-1}$$

*for all $j \geqslant 0$. $C_0 = 2r$, $C_1 = 128 \left( \frac{\kappa_1 M}{r \rho^\alpha} \right)^2$ with $\alpha, \kappa_1$ as defined in Lemma 4.1.2.*

*Proof modified from Hairer et al. [26].* By Equation 4.5, $S_j$ solves the differential equation from Lemma 4.1.2 for $g = K_j$, where $K_j$ is defined in Equation 4.6. Fix some $N$ and let $\delta_1 = \frac{\rho}{4}, \delta_i = \frac{\rho}{4N}$ for all $i > 1$. Abbreviate $\|K_k\|_i = \|K_k(\vec{J}^*, \cdot)\|_{\rho - \sum_{j=1}^i \delta_j}$ for all $i$. Lemma 4.1.2 now implies that

$$\left\| \frac{\partial S_j}{\partial \vec{\varphi}} \right\|_j \leqslant \kappa_1 \delta_j^{-\alpha} \|K_j\|_{j-1}.$$

We need to bound the right hand side.

Cauchy's estimate tells us that

$$\left| \frac{1}{i!} \frac{\partial^i H_0}{\partial \vec{I}^i}(v_1, \ldots, v_i) \right| \leqslant \frac{M}{r^i} |v_1| \ldots |v_i|$$

so we have

$$\|K_j\|_{j-1} \leqslant \sum_{i=2}^{j} \sum_{k_1 + \ldots + k_i = j} \frac{M}{r^i} \left\| \frac{\partial S_{k_1}}{\partial \vec{\varphi}} \right\|_{k_1} \cdots \left\| \frac{\partial S_{k_i}}{\partial \vec{\varphi}} \right\|_{k_i} + \sum_{i=1}^{j-1} \sum_{k_1 + \ldots + k_i = j-1} \frac{M}{r^i} \left\| \frac{\partial S_{k_1}}{\partial \vec{\varphi}} \right\|_{k_1} \cdots \left\| \frac{\partial S_{k_i}}{\partial \vec{\varphi}} \right\|_{k_i}.$$

Now we can combine the two bounds to get a recursive bound. Define $\mu_j = \frac{M\kappa_1}{r \delta_j^\alpha}$ and we have

$$\frac{1}{r} \left\| \frac{\partial S_j}{\partial \vec{\varphi}} \right\|_j \leqslant \beta_j$$

where $\beta_1 = \mu_1$ and

$$\beta_j = \mu_j \sum_{i=2}^{j} \sum_{k_1 + \ldots + k_i = j} \beta_{k_1} \ldots \beta_{k_i} + \mu_j \sum_{i=1}^{j-1} \sum_{k_1 + \ldots + k_i = j-1} \beta_{k_1} \ldots \beta_{k_i}.$$

To tackle this, define the generating function $b(\zeta) = \sum_{j=1}^{\infty} \beta_j \zeta^j$ and we then have

$$b(\zeta) - \mu_1 \zeta = \mu_N \left( \frac{1}{1 - b(\zeta)} - 1 - b(\zeta) \right) + \mu_N \zeta \left( \frac{1}{1 - b(\zeta)} - 1 \right).$$

This is a quadratic equation; it can be solved to yield

$$b(\zeta) = \frac{1}{2} \frac{1 + (\mu_1 - \mu_N)\zeta}{1 + \mu_N} - \sqrt{\frac{1}{4} \left( \frac{1 + (\mu_1 - \mu_N)\zeta}{1 + \mu_N} \right)^2 - \frac{\mu_1}{1 + \mu_N} \zeta} \leqslant \frac{1 + (\mu_1 - \mu_N)\zeta_{\max}}{2(1 + \mu_N)}.$$

Picking $\zeta_{\max} = \frac{1}{C_1 N^\alpha}$, we can verify that $b(\zeta)$ is analytic in the region $|\zeta| \leqslant \zeta_{\max}$ (the quantity under the square root is positive). By Cauchy's estimate, we now have

$$\left\|\frac{\partial S_j}{\partial \vec{\varphi}}\right\|_N \leqslant r\beta_N \leqslant 2r(C_1 N^\alpha)^{N-1} \leqslant C_0(C_1 N^\alpha)^{N-1}$$

as desired. $\qquad\square$

The last thing we need is a bound on the remainder terms.

**Lemma 4.1.4.** *Under the same conditions as in Lemma 4.1.3, with $r \leqslant 1, C_1 N^\alpha \leqslant \frac{1}{2\epsilon}$, we have*

$$\|R_N(\vec{J}^*, \cdot)\|_{\frac{\varrho}{2}} \leqslant 4Mr\left(\frac{4C_1}{r}N^\alpha\right)^N.$$

*Proof from Hairer et al. [26].* By Equations 4.7 and 4.6, the remainder terms look something like

$$\frac{1}{i!}\frac{\partial^i H_{k_0}}{\partial \vec{I}^i}(Q_{k_1}, \dots, Q_{k_i})$$

with

$$Q_k = \frac{\partial S_k}{\partial \vec{\varphi}} + \epsilon \frac{\partial S_{k+1}}{\partial \vec{\varphi}} + \dots + \epsilon^{N-k-1}\frac{\partial S_{N-1}}{\partial \vec{\varphi}}.$$

This is just everything that's left over in our Taylor expansion after we take terms of order $N - 1$ or lower.

By Lemma 4.1.3, we have

$$\|Q_k(\vec{J}^*, \cdot)\|_{\frac{\varrho}{2}} \leqslant \sum_{j=k}^{N-1} \epsilon^{j-k}C_0(C_1 j^\alpha)^j \leqslant C_0 \sum_{j=1}^{N-1} 2^{-(j-k)}\left(\frac{j}{N}\right)^{\alpha j}(C_1 N^\alpha)^k \leqslant 2C_0(C_1 N^\alpha)^k$$

using our condition and the formula for a geometric series. Cauchy's estimate now yields

$$\left\|\frac{1}{i!}\frac{\partial^i H_{k_0}}{\partial \vec{I}^i}(Q_{k_1}, \dots, Q_{k_i})\right\|_{\frac{\varrho}{2}} \leqslant \frac{M}{r^i}2C_0(C_1 N^\alpha)^N \leqslant 4Mr\left(\frac{C_1}{r}N^\alpha\right)^N.$$

A stars-and-bars-style counting argument implies that there are $2\binom{N+i-1}{i}$ terms of this form for a given $i$. We may thus bound the total number of terms independent of $i$ by

$$2\sum_{i=0}^{N-1}\binom{N+i-1}{i} \leqslant \sum_{i=0}^{2N-1}\binom{2N-1}{i} = 4^N.$$

This gives us the desired estimate. $\qquad\square$

## 4.1.3 Invariant Tori under Perturbations

With these bounds computed, we can understand how closely invariant tori are maintained under perturbations. Define

$$\vec{\omega}_{\epsilon,N}(\vec{J}) = \frac{\partial}{\partial \vec{J}} K(\vec{J}, \vec{\theta}) \tag{4.8}$$

where $K$ is the truncated Hamiltonian to $N$ terms in the new coordinates $(\vec{J}, \vec{\theta})$. This is thus the expected angular velocity of the angular coordinates $\vec{\theta}$ if our perturbed Hamiltonian was actually completely integrable.

**Theorem 4.1.5.** *Let $H_0$ be real-analytic on $B(\vec{J}^*, r)$ for $\vec{J}^* \in \mathbb{R}^n$ and let $H_1$ be real-analytic on $B(\vec{J}^*, r) \times U_\rho$ with $r \leqslant 1, \rho \leqslant 1$. Suppose that $\omega(\vec{J}^*) = \frac{\partial H_0}{\partial \vec{I}}(\vec{J}^*)$ satisfies Siegel's diophantine condition (Equation 3.3). Then there exist constants $\epsilon_0, c_0, C > 0$ such that for every $0 < \beta \leqslant 1$ and $\epsilon \leqslant \epsilon_0$, we have a real-analytic symplectic transformation $(\vec{I}, \vec{\varphi}) \mapsto (\vec{J}, \vec{\theta})$ such that trajectories $(\vec{J}(t), \vec{\varphi}(t))$ starting at $||\vec{J}(0) - \vec{J}^*|| \leqslant c_0 \epsilon^{2\beta}$ satisfy*

$$||\vec{J}(t) - \vec{J}(0)|| \leqslant Ct \exp\left(-c\epsilon^{-\frac{\beta}{\alpha}}\right)$$

*and*

$$||\vec{\theta}(t) - \vec{\omega}_{\epsilon,N}(\vec{J}(0))t - \vec{\theta}(0)|| \leqslant Ct\epsilon^{-2\beta} \exp\left(-c\epsilon^{-\frac{\beta}{\alpha}}\right)$$

*for $t \leqslant \exp\left(\frac{1}{2}c\epsilon^{-\frac{\beta}{\alpha}}\right)$. Here, $\alpha = \nu + n + 1, c = \left(\frac{16C_1 e}{r}\right)^{-\frac{1}{\alpha}}$ with $C_1$ from Lemma 4.1.3. Further, for $(\vec{I}, \vec{\varphi}) \mapsto (\vec{J}, \vec{\theta})$, we have*

$$||\vec{I} - \vec{J}|| \leqslant C\epsilon$$

*for $||\vec{J} - \vec{J}^*|| \leqslant c_0 \epsilon^{2\beta}, \vec{\theta} \in U_{\frac{\rho}{2}}$.*

Intuitively, this theorem tells us that invariant tori under perturbation become near-invariant tori, i.e. they are close to invariant (within exponentially decaying error) for exponentially long times. We also see that the transformations are close to the identity.

*Proof from Hairer et al. [26].* If $H_1(\vec{J}, \vec{\varphi})$ is a trigonometric polynomial in $\vec{\varphi}$ of degree $m$, then by construction $K_j$ and $S_j$ are trigonometric polynomials of degree $jm$. We have

$$|k \cdot \omega(\vec{J})| \geqslant |k \cdot \omega(\vec{J}^*)| - |k|(\max ||\omega'||)||\vec{J} - \vec{J}^*||$$

by the Mean Value Theorem, so there exists $\delta > 0$ such that

$$|k \cdot \omega(\vec{J})| \geqslant \frac{1}{2}\gamma|k|^{-\nu}$$

for $||\vec{J} - \vec{J}^*|| \leqslant \delta, |k| \leqslant jm$. This $\delta$ must be proportional to $\gamma(jm)^{-\nu-1}$; we may let $\delta$ be proportional to $\gamma(jm)^{-\alpha}$ as that is weaker. Our estimate for the remainder term and generating function from Lemmas 4.1.3 and 4.1.4 then hold on this region.

For general $H_1$, we may approximate it by a trigonometric polynomial of degree $m$ with error $\mathcal{O}\left(e^{-m\frac{\rho}{2}}\right)$ on $B(\vec{J}^*,r) \times U_\rho$. Pick $m = \frac{2N}{\rho}$; then this error is just $\mathcal{O}\left(e^{-N}\right)$. Then our bounds for the remainder term and generating function still hold.

Thus, for some constant $c^*$ and $C_2 = \frac{16C_1}{r}$, we have

$$\left\|\frac{\partial S_j}{\partial \vec{\varphi}}(\vec{J},\vec{\varphi})\right\| \leqslant C_0(4C_1 j^\alpha)^{j-1}$$

for $\|\vec{J} - \vec{J}^*\| \leqslant c^*(jm)^{-\alpha}, \vec{\varphi} \in U_{\frac{\rho}{2}}$ and

$$|R_N(\vec{J},\vec{\varphi})| \leqslant 4Mr(C_2 N^\alpha)^N$$

for $\|\vec{J} - \vec{J}^*\| \leqslant c^*(Nm)^{-\alpha}, \vec{\varphi} \in U_{\frac{\rho}{2}}$. Pick $c_0$ so these thresholds are satisfied.

The Hamiltonian equations for our system must be

$$\dot{\vec{J}} = -\frac{\partial K}{\partial \vec{\theta}}(\vec{J},\vec{\theta}) = -\epsilon^N \frac{\partial R_N}{\partial \vec{\varphi}} \frac{\partial \vec{\varphi}}{\partial \vec{\theta}} = \mathcal{O}\left(\epsilon^N (C_2 N^\alpha)^N\right)$$

and

$$\dot{\vec{\theta}} = \frac{\partial K}{\partial \vec{J}}(\vec{J},\vec{\theta}) = \vec{\omega}_{\epsilon,N}(\vec{J}) + \mathcal{O}\left((Nm)^\alpha \epsilon^N (C_2 N^\alpha)^N\right).$$

Pick $N$ such that $C_2 N^\alpha = \frac{1}{e\epsilon^\beta}$ so we then get

$$\dot{\vec{J}} = \mathcal{O}\left(\exp\left(-c\epsilon^{-\frac{\beta}{\alpha}}\right)\right), \dot{\vec{\theta}} = \vec{\omega}_{\epsilon,N}(\vec{J}) + \mathcal{O}\left(\epsilon^{-2\beta} \exp\left(-c\epsilon^{-\frac{\beta}{\alpha}}\right)\right).$$

for $\|\vec{J} - \vec{J}^*\| \leqslant c_0 \epsilon^{2\beta}$ and $c = (C_2 e)^{-\frac{1}{\alpha}}$, which is the first desired statement.

For the second part, recall that $\vec{I} = \frac{\partial S}{\partial \vec{\varphi}} = \vec{J} + \epsilon^j \sum_{j=1}^{N} \frac{\partial S_j}{\partial \vec{\varphi}}$ so on the given range we have

$$\|\vec{I} - \vec{J}\| \leqslant \sum_{j=1}^{N} \epsilon^j C_0(4C_1 j^\alpha)^{j-1}$$

which is clearly an $\mathcal{O}(\epsilon)$ bound for sufficiently small $\epsilon$; we can adjust the threshold $\epsilon_0$ as needed to ensure that this is true. $\qquad \square$

## 4.2 Applying Perturbation Theory to Störmer-Verlet

Now that we have the necessary building blocks, our final step is to apply perturbation theory to the Störmer-Verlet method and bound the resulting errors; we will use this to bound errors in energy conservation and time discretization error. Our general strategy for applying the perturbation theory of completely integrable systems to analyzing the Störmer-Verlet method is relatively simple. We know by Theorem 2.2.1 and by Theorem 2.2.2 that a modified Hamiltonian exists that describes dynamics according to the Störmer-Verlet process. Further, since the Störmer-Verlet method is second-order, the leading-order correction term has order $h^2$. Thus we may pick $\epsilon = h^2$ and define our perturbation using this modified Hamiltonian.

We'll begin by applying these methods to refine the estimate we made for energy conservation under Störmer-Verlet in Theorem 2.2.3.

## 4.2.1 Energy Conservation

Complete integrability allows us to improve bounds on conservation of energy by removing the linear error growth in time we found in Theorem 2.2.3.

**Theorem 4.2.1.** *Apply Störmer-Verlet to a completely integrable Hamiltonian system. Suppose that $\omega(\vec{I}^*)$ satisfies Siegel's diophantine condition. Then there exist constants $C, c, h_0$ such that for all step sizes $h \leqslant h_0$, every numerical solution starting at $\vec{I}_0$ with $||\vec{I}_0 - \vec{I}^*|| \leqslant c |\log h|^{-\nu-1}$ satisfies*

$$||\vec{I}_n - \vec{I}_0|| \leqslant Ch^2$$

*for $t = nh \leqslant h^{-2}$.*

*Proof modified from Hairer et al. [26].* By Theorem 2.2.2, we know there exists a modified Hamiltonian

$$\widetilde{H}(\vec{p}, \vec{q}) = H(\vec{p}, \vec{q}) + h^2 H_3(\vec{p}, \vec{q}) + h^3 H_4(\vec{p}, \vec{q}) + h^4 H_5(\vec{p}, \vec{q})$$

where we truncate at the $h^4$ term. This modified Hamiltonian describes dynamics under Störmer-Verlet up to $\mathcal{O}(h^5)$. Picking $\epsilon = h^2$, we now have a perturbed Hamiltonian in action-angle coordinates. Let $\vec{J}_n$ (or $\vec{I}_n$ in the original coordinates) be the numerical trajectory and $\vec{J}(t)$ be the trajectory along the truncated differential equation.

   We now apply Theorem 4.1.5, picking $h_0$ from the given value of $\epsilon_0$. This tells us that over (significantly more than) the relevant time range we have

$$|\vec{J}(t) - \vec{J}_0|| \leqslant Ct \exp\left(-ch^{-\frac{2\beta}{\alpha}}\right), ||\vec{I}_0 - \vec{J}_0||, ||\vec{I}_n - \vec{J}_n|| \leqslant Ch^2.$$

Note that the first bound is smaller than the second one over the given time range $t \leqslant h^{-2}$. All we need is a bound on $||\vec{I}_n - \vec{J}_0||$.

   One step of the numerical method must look like $\vec{J}_{n+1} = \vec{J}_n + \mathcal{O}(h^5)$ by the construction of the truncated form. Thus at $t = nh$ we have $\vec{J}_n = \vec{J}(t) + \mathcal{O}(th^4)$ and we have $t \leqslant h^{-2}$, so that last term is $\mathcal{O}(h^2)$. Adding together the bounds that we have provided yields that $||\vec{I}_n - \vec{I}_0|| \leqslant Ch^2$ as desired. $\qquad\square$

   Since the first coordinate of the original $\vec{I}$ action-angle coordinates is the unperturbed Hamiltonian, this is a bound on the energy of the system under simulation that does not grow with time. Thus this is a significant improvement over the bounds provided in Theorem 2.2.3.

   This estimate can be significantly improved to be accurate over exponential times and starting points of polynomial distance away, just as with Theorem 4.1.5 [26]. That requires a fancier estimate for $||\vec{J}_n - \vec{J}(t)||$ relying on analyticity; the details can be found in Hairer et al. [26]. We will not need this bound for our work in understanding time discretization error.

## 4.2.2 Bounding Time Discretization Error

We will now simultaneously bound time discretization and sampling error. We present this result with the use of a filter function; the modification to the typical case without a filter function is obvious.

Given a filter function $\xi$, define the time average

$$\langle A \rangle_\xi^{\text{Rie}}(T) = \frac{\sum_{j=0}^{n-1} \xi\left(\frac{j}{n}\right) A(\vec{p}_j, \vec{q}_j)}{\sum_{j=0}^{n-1} \xi\left(\frac{j}{n}\right)} \tag{4.9}$$

where $\{\vec{p}_j, \vec{q}_j\}$ represent the trajectory computed via Störmer-Verlet and $n$ is the number of time steps taken. Let $h$ be the step size as usual. This is the discretized version of the filter-function time average we computed earlier. We want to bound the difference between this and $\langle A \rangle$.

**Theorem 4.2.2.** *Consider a completely integrable Hamiltonian system with real-analytic Hamiltonian H simulated under Störmer-Verlet. Suppose our filter function is $(d+1)$ times differentiable with $\xi^{(j)}(0) = \xi^{(j)}(1) = 0$ for all $0 \leqslant j \leqslant d-1$ and our observable A is real-analytic. In action-angle coordinates, suppose $\vec{\omega}(\vec{I}^*)$ satisfies Siegel's diophantine condition (Equation 3.3). Then there exist positive constants $h_0, c_0, c_1, C$ such that for all trajectories starting at $||\vec{I}_0 - \vec{I}^*|| \leqslant c_0 h^{2\mu}$ where $\mu \leqslant \min(2, \nu + n + 1)$ and the step size $h \leqslant h_0$, we have for times $T = nh \leqslant \exp\left(c_1 h^{-\frac{\mu}{\alpha}}\right)$ that*

$$|\langle A \rangle_\xi^{\text{Rie}}(T) - \langle A \rangle| \leqslant C\left(\frac{1}{T^{d+1}} + h^2\right).$$

Note that the $\mathcal{O}\left(\frac{1}{T^{d+1}}\right)$ error bound corresponds to sampling error, so the only additional contribution from time discretization error is a $\mathcal{O}\left(h^2\right)$ bound. Thus for exponentially long times, time discretization error does not increase the error in our measured observable.

*Proof from Cances et al. [9].* The conditions provided are sufficient to apply Theorem 4.1.5 with $\epsilon = h^2$, as above. Switch to the $(\vec{J}, \vec{\theta})$ coordinate system defined using that theorem. We then have using Equation 4.9

$$\langle A \rangle_\xi^{\text{Rie}}(T) = \frac{\sum_{j=0}^{n-1} \xi\left(\frac{j}{n}\right) A(\vec{J}_j, \vec{\theta}_j)}{n \sum_{j=0}^{n-1} \frac{1}{n} \xi\left(\frac{j}{n}\right)}.$$

We now take a Fourier transform and apply the trajectories from Theorem 4.1.5. Define $\vec{\omega}_h = \vec{\omega}_{h^2,5}(\vec{J}(0))$ (recalling the notation from Equation 4.8). Expanding our Fourier transform, we get

$$\langle A \rangle_\xi^{\text{Rie}}(T) = \frac{1}{n \sum_{j=0}^{n-1} \frac{1}{n} \xi\left(\frac{j}{n}\right)} \sum_{\vec{k} \in \mathbb{Z}^n} \hat{A}(\vec{J}_0, \vec{k}) e^{i\vec{k} \cdot \vec{\theta}_0} \left(\sum_{j=0}^{n-1} \xi\left(\frac{j}{n}\right) e^{ijh(\vec{k} \cdot \vec{\omega}_h)}\right) + \mathcal{O}\left(\exp\left(-ch^{-\frac{\mu}{\alpha}}\right)\right). \tag{4.10}$$

Theorem 4.1.5 tells us that our transformation is $\mathcal{O}(h^2)$ away from the identity, so as before we have

$$\hat{A}(\vec{J}_0, \vec{0}) = \langle A \rangle + \mathcal{O}(h^2).$$

We now need to get the $\mathcal{O}\left(\frac{1}{T^{k+1}}\right)$ bound corresponding to sampling error. We will proceed by rearranging the above equation's nonzero terms into something we can more easily apply Siegel's diophantine condition to as before. First, we have the following motivating lemma:

**Lemma 4.2.3.** *For a filter function $\xi$ satisfying the above conditions, given a complex number $z \neq 1, |z| = 1$, we have*

$$\left| \sum_{j=0}^{n-1} \xi\left(\frac{j}{n}\right) z^j \right| \leqslant \frac{2e^2 d^d}{n^d |1-z|^{d+1}} \left( |\xi^{(d)}(0)| + |\xi^{(d)}(1)| + ||\xi^{(d+1)}||_{L^\infty} \right)$$

*assuming $d < n$.*

The proof of Lemma 4.2.3 is by expansion and the Mean Value Theorem; it may be found in Cances et al. [9]. We'd like to apply Lemma 4.2.3 with $z = e^{ijh(\vec{k} \cdot \vec{\omega}_h)}$, but we must make sure that Siegel's diophantine condition is satisfied for $\vec{\omega}_h$, else the RHS of Lemma 4.2.3 will be unbounded and thus useless. We currently only have Siegel's diophantine condition for $\vec{\omega}(\vec{I}^*)$, but fortunately, Siegel's diophantine condition holds in a modified form for most $\vec{k}$ at $\vec{\omega}_h$; we now prove this fact.

**Lemma 4.2.4.** *For all $h \leqslant h_0$ and all $\vec{k} \in \mathbb{Z}^n$, we have*

$$|\vec{k} \cdot (h\vec{\omega}_h)| \leqslant \frac{\gamma}{2} |\vec{k}|^{-\nu} \implies |\vec{k}| \geqslant ch^{-\frac{2}{\nu+1}}$$

*for some positive constant $c$.*

*Proof from Cances et al. [9].* We have by Theorem 4.1.5 that $h\vec{\omega}_h = \vec{\omega} + \mathcal{O}\left(|h\vec{\omega}|^2\right)$ so for some constant $C$ we have

$$\frac{\gamma}{2} |\vec{k}|^{-\nu} \geqslant |\vec{\omega} \cdot \vec{k}| - C|\vec{k}||h\vec{\omega}|^2 \geqslant \gamma |\vec{k}|^{-\nu} - C|\vec{k}||h\vec{\omega}|^2$$

implying that

$$|\vec{k}| \geqslant \left( \frac{\gamma}{2C|\vec{\omega}|^2} h^{-2} \right)^{\frac{1}{\nu+1}}$$

so we may pick $c = \frac{\gamma}{2C|\vec{\omega}|^2}$. $\qquad\qquad\qquad\qquad\qquad\qquad\square$

By Lemma 4.2.4, all $\vec{k}$ with $|\vec{k}| < ch^{-\frac{2}{\nu+1}}$ satisfy Siegel's diophantine condition with $h\vec{\omega}_h$ and a modified constant $\frac{\gamma}{2}$. We can therefore split the summation in Equation 4.10

and bound as follows:

$$
|\langle A \rangle_{\xi}^{\mathrm{Rie}}(T) - \hat{A}(\vec{J}_0, \vec{0})| \leqslant \frac{1}{n \sum_{j=0}^{n-1} \frac{1}{n} \xi\left(\frac{j}{n}\right)} \sum_{\vec{k} \in \mathbb{Z}^n, 0 < |\vec{k}| < ch^{-\frac{2}{\nu+1}}} |\hat{A}(\vec{J}_0, \vec{k})| \left| \sum_{j=0}^{n-1} \xi\left(\frac{j}{n}\right) e^{ijh(\vec{k} \cdot \vec{\omega}_h)} \right|
$$

$$
+ \sum_{\vec{k} \in \mathbb{Z}^n, |\vec{k}| \geqslant ch^{-\frac{2}{\nu+1}}} |\hat{A}(\vec{J}_0, \vec{k})|
$$

$$
\leqslant \frac{1}{n^{d+1} \sum_{j=0}^{n-1} \frac{1}{n} \xi\left(\frac{j}{n}\right)} \sum_{\vec{k} \in \mathbb{Z}^n, 0 < |\vec{k}| < ch^{-\frac{2}{\nu+1}}} \frac{|\hat{A}(\vec{J}_0, \vec{k})|}{\left|1 - e^{ih(\vec{k} \cdot \vec{\omega}_h)}\right|^{d+1}} c_0
$$

$$
+ \sum_{\vec{k} \in \mathbb{Z}^n, |\vec{k}| \geqslant ch^{-\frac{2}{\nu+1}}} |\hat{A}(\vec{J}_0, \vec{k})|
$$

by Lemma 4.2.3 for some $c_0 > 0$.

We now have $\frac{1}{|1 - e^{i\vec{k} \cdot h\vec{\omega}}|} \leqslant \frac{c_1}{h|\vec{k} \cdot \vec{\omega}|}$ by Taylor expansion, so we may estimate the first term above as

$$
\frac{c_0 c_1^{d+1}}{T^{d+1} \sum_{j=0}^{n-1} \frac{1}{n} \xi\left(\frac{j}{n}\right)} \sum_{\vec{k} \in \mathbb{Z}^n, 0 < |\vec{k}| < ch^{-\frac{2}{\nu+1}}} \frac{|\hat{A}(\vec{J}_0, \vec{k})|}{|\vec{k} \cdot \vec{\omega}_h|^{d+1}}.
$$

We have previously shown in the proof of Theorem 3.2.3 that the sum is an $\mathcal{O}\left(\frac{1}{T^{d+1}}\right)$ quantity and $\sum_{j=0}^{n-1} \frac{1}{n} \xi\left(\frac{j}{n}\right) \to \int_0^1 \xi(x)\,dx$ and can thus be also bounded by a constant, so this sum is $\mathcal{O}\left(\frac{1}{T^{d+1}}\right)$.

The second sum above is just the sum of exponentially decaying Fourier coefficients, which decays as $\mathcal{O}\left(\exp\left(h^{-\frac{2}{\nu+1}}\right)\right)$, clearly dominated by our previous $\mathcal{O}(h^2)$ error estimate. Combining these errors, we get the desired statement. $\qquad\square$

As before, we may pick $\xi(x) = e^{-\frac{1}{x(1-x)}}$ to replace the $\frac{1}{T^{d+1}}$ term by an exponentially decaying bound in $T$.

The results shown here demonstrate that time discretization error is bounded by $\mathcal{O}(h^2)$ when using the Störmer-Verlet method. Most importantly, for exponentially long times this bound is constant, implying that our time discretization error should not increase over a reasonable timeframe associated with an MD simulation.

# Chapter 5

# Quantum Mechanics and Force-Field Error

We now turn to force-field error, generally considered the most significant source of error in MD simulations [35]. Unfortunately, our understanding of force-field error, both bounding the error associated with well-known force fields and determining which force field would minimize such error, is still rather poor. In this chapter, we present some recent results on hypothetical optimal bounds of force-field error.

Generally speaking, force-field error occurs because atoms move according to the rules of quantum mechanics, not classical mechanics, and thus do not exactly follow Newton's second law. To better understand force-field error, we will therefore begin with a brief review of quantum mechanics from a mathematical perspective. Here, we will only present the facts cogent for our discussion of force fields; for a more thorough introduction to quantum mechanics, see Griffiths [22].

In what follows, we assume we are in the non-relativistic regime; relativistic effects are usually insignificant in chemistry. Following the more typical custom in quantum mechanics, we use $\vec{x}$ instead of $\vec{q}$ for positions.

## 5.1 Introduction to Quantum Mechanics

### 5.1.1 The State Space and the Schrödinger Equation

The quantum-mechanical state of an $N$-particle system is given by a complex-valued **wavefunction** $\psi(\vec{x}, t)$ where $\vec{x} \in \mathbb{R}^{3N}$ represents the position of all $N$ particles and $t \in \mathbb{R}$ is time. $|\psi(\vec{x}, t)|^2$ is the probability distribution of the $N$ particle's positions; we'll discuss how to compute probability distributions of other observables later. Since this probability distribution must be normalizable, we require

$$\int_{\mathbb{R}^{3N}} |\psi(\vec{x}, t)|^2 \, d\vec{x} < \infty \tag{5.1}$$

at any given time $t$. Thus our wavefunction at any given time $t$ belongs to the Hilbert space $L^2(\mathbb{R}^{3N})$. The inner product on this vector space is denoted

$$\langle \psi(t) | \phi(t) \rangle := \int_{\mathbb{R}^{3N}} \overline{\psi}(\vec{x}, t) \phi(\vec{x}, t) \, dx$$

at a given time $t$ [23].

There are many ways to motivate the Schrödinger Equation, either by analogy to classical mechanics or by comparison with experimental results. Either way, we get the time-dependent Schödinger equation (TDSE)

$$i\hbar \frac{\partial}{\partial t} \psi(\vec{x}, t) = H\psi(\vec{x}, t) \tag{5.2}$$

where $\hbar$ is Planck's constant and $H$ is an operator on our Hilbert space defined by

$$H\psi(\vec{x}, t) = -\sum_{i=1}^{N} \frac{\hbar^2}{2m_i} \Delta_i \psi + V\psi. \tag{5.3}$$

Here, $m_i$ and $\Delta_i$ are the mass and Laplacian associated with the $i$th particle and $V$ is the exact potential. Essentially all potentials we will deal with will be electromagnetic.

We would like to solve this equation given some initial condition $\psi|_{t=0} = \psi_0$ to find $\psi$ at all times. Our solution must not only solve the Schrödinger equation but also conserve probability, or continue to be a probability distribution at all times. Thus we require that $\langle \psi(t) | \psi(t) \rangle$ is constant.

**Theorem 5.1.1.** *Probability is conserved, or $\langle \psi(t) | \psi(t) \rangle$ is constant, if $\psi$ satisfies the Schrödinger equation and $H$ is **symmetric**, i.e.*

$$\langle H\varphi | \phi \rangle = \langle \phi | H\varphi \rangle$$

*for all functions $\phi, \varphi \in L^2(\mathbb{R}^{3N})$.*

*Proof from Gustafson and Sigal [23].* Assuming that $\psi$ satisfies the Schrödinger equation, we have

$$\frac{d}{dt} \langle \psi(t) | \psi(t) \rangle = \langle \dot{\psi}(t) | \psi(t) \rangle + \langle \psi(t) | \dot{\psi}(t) \rangle = \left\langle \frac{1}{i\hbar} H\psi(t) \,\middle|\, \psi(t) \right\rangle + \left\langle \psi(t) \,\middle|\, \frac{1}{i\hbar} H\psi(t) \right\rangle$$

which is zero for all times $t$ iff $\langle H\phi | \phi \rangle = \langle \phi | H\phi \rangle$ for all $\phi$ (since $\psi(0)$ can be anything we like). This is true iff $H$ is symmetric by standard arguments using the polarization identity. $\qquad\square$

The other condition we need for dynamics to exist, i.e. for the TDSE to have a solution, is for our Hamiltonian to be **self-adjoint**. To prove this, we assume standard results about the existence of exponentials of operators; for their proofs, see [23].

**Theorem 5.1.2.** *Dynamics exist iff $H$ is self-adjoint.*

*Proof Outline from Gustafson and Sigal [23].* Recall the exponential operator $e^A = \sum_{n=0}^{\infty} \frac{A^n}{n!}$. This is well-defined in general using approximations of $A$ by bounded operators [23]. Further, if $A$ is self-adjoint, then $e^{iA}$ is **unitary**, where an operator $U$ is unitary if $UU^* = U^*U = $ id. Thus, if $H$ is self-adjoint, then $U(t) := e^{-\frac{i}{\hbar}Ht}$ exists and is unitary. We may now define $\psi(t) = U(t)\psi_0$ for our initial conditions $\psi_0$. A straightforward computation shows that

$$i\hbar \frac{\partial}{\partial t} e^{-\frac{i}{\hbar}Ht} \psi_0 = H e^{-\frac{i}{\hbar}Ht} \psi_0$$

so this solution satisfies the Schrödinger equation. This can also be checked to conserve probability. $\qquad\square$

Thus we will require that our Hamiltonians be self-adjoint. The Hamiltonian describing molecular systems does satisfy this property.

### 5.1.2 Observables

Now that we understand dynamical properties of our system, we need to talk about how to measure observables. **Observables** in quantum mechanics correspond to self-adjoint operators.

**Example 5.1.3.** Examples of observables:

- The position operator $x_j$ in the $j$th coordinate: $\psi \mapsto x_j\psi$.

- The momentum operator $p_j$ in the $j$th coordinate: $\psi \mapsto -i\hbar\frac{\partial}{\partial x_j}\psi$.

- The Hamiltonian operator $H$.

- The angular momentum operator $L_j = (x \times p)_j$.

The commutator of any two operators is defined as $[A, B] = AB - BA$. For example,

$$[x_j, p_j] = i\hbar. \tag{5.4}$$

This is the canonical commutation relation and equivalent to the standard uncertainty principle [23].

We can compute the expectation value of any observable by computing

$$\langle A \rangle = \frac{\langle \psi | A\psi \rangle}{\langle \psi | \psi \rangle}.$$

This corresponds to a weighted integral over a probability distribution; the essential difference between classical and quantum mechanics is that our probability "distribution" may now be complex-valued [23].

The time evolution of this expectation value may be computed using the Heisenberg equation

$$\frac{d}{dt}\langle A \rangle = \left\langle \psi \left| \frac{i}{\hbar}[H, A]\psi \right. \right\rangle. \tag{5.5}$$

Applying Equation 5.5 to $x_j, p_j$ and using Equation 5.4 yields the system

$$m\frac{d}{dt}\langle x_j \rangle = \langle p_j \rangle, \frac{d}{dt}\langle p_j \rangle = \left\langle -\frac{d}{dx_j}V \right\rangle$$

which is the quantum-mechanical analog of Hamilton's Equations (equations 2.5) [23].

### 5.1.3 Spectra of Operators

The most efficient and standard way of solving the Schrödinger Equation is by understanding the spectrum of the Hamiltonian. Here we review some results from operator spectral theory before describing how understanding the spectrum of the Hamiltonian helps us solve the Schrödinger Equation. We omit the proofs of many of these results as they are tangential to our discussion.

We will for now specialize to a one-particle system and later generalize to multi-particle systems.

**Definition 5.1.4.** The **spectrum** of an operator $A$ on a Hilbert space $\mathcal{H}$ is the set $\sigma(A)$ of $\lambda \in \mathbb{C}$ such that $A - \lambda$ is not invertible, i.e. has no bounded inverse.

The complement of the spectrum is called the resolvent set of $A$ $\rho(A)$ and for $\lambda \in \rho(A)$, $(A - \lambda)^{-1}$ is the resolvent of $A$.

The first result we need is a better characterization of the spectrum of a self-adjoint operator:

**Theorem 5.1.5** (Weyl). *The spectrum of $A$ consists of $\lambda$ such that:*

1. $(A - \lambda)\psi = 0$ *for some $\psi \in \mathcal{H}$. Then $\lambda$ is an eigenvalue, $\psi$ is an eigenvector, and $\lambda$ is in the* ***point spectrum*** *or* ***discrete spectrum***. *Further, $\lambda$ is isolated and has finite multiplicity.*

2. *There exists a* ***Weyl sequence*** *$\{\psi_n\} \subset \mathcal{H}$, or a sequence where*

$$||\psi_n|| = 1, ||(A - \lambda)\psi_n|| \to 0$$

*as $n \to \infty$, and $\langle \phi | \psi_n \rangle \to 0$ for all $\phi \in \mathcal{H}$ as $n \to \infty$. In this case, $\lambda$ is in the* ***continuous spectrum*** *or* ***essential spectrum***.

A proof may be found in Hislop and Sigal [29]. Note further that the spectrum of a self-adjoint operator is real.

**Example 5.1.6.** We can compute the spectra of some simple operators here in $L^2(\mathbb{R}^{3N})$:

1. $\sigma(x_j) = \sigma_c(x_j) = \mathbb{R}$. It's easy to see that $x_j$ has no eigenvalues. For any $\lambda \in \mathbb{R}$, we can construct a Weyl sequence for $x_j$ and $\lambda$. Work in one dimension for now (ignoring all the others). Pick $\psi_n(x) = \sqrt{n}\phi(n(x - \lambda))$ for $\phi$ any fixed non-negative function on $[-1, 1]$ such that $\int |\phi|^2 = 1$. It's easy to verify that this is in fact a Weyl sequence, so we have the desired [23].

2. $\sigma(p_j) = \sigma_c(p_j) = \mathbb{R}$. We could construct another set of Weyl sequences, but just note that $p_j$ and $x_j$ are conjugates by the Fourier transform, which is unitary, and conjugating by a unitary transformation preserves the spectrum [23].

3. Given continuous $V : \mathbb{R}^3 \to \mathbb{C}$, the spectrum of the corresponding multiplication operator is the closure of range($V$). The same Weyl sequence as in the first part works here as well.

4. $\sigma(-\Delta) = \sigma_c(-\Delta) = [0, \infty)$. This is equivalent to solving the Helmholtz equation.

Once we can compute the spectra of an operator, we may then compute its probability distribution. Given an operator $A$, we can compute eigenstates of that operator $\psi_x$ for all elements of the spectrum $x \in \sigma(A)$ and expand any particular wavefunction in terms of those eigenstates

$$\psi = \int_{\sigma(A)} c(x)\, \psi_x$$

for some coefficients $c(x)$. Then the probability distribution function is defined to be $|c(x)|^2$, assuming that our wavefunction is normalized initially. We can see that this does get the desired expectation value noted previously.

We will need a slightly stronger criterion to find the spectrum of the Hamiltonian, which holds for Schrödinger Hamiltonians.

**Theorem 5.1.7.** *Given $H = -\frac{\hbar^2}{2m}\Delta + V$ with real potential $V(\vec{x})$ continuous and bounded from below, then $\sigma_c(H)$ consists of $\lambda$ for which there exists a **spreading sequence** $\{\psi_n\}$. A spreading sequence is identical to a Weyl sequence, but instead of the third condition above we require that for any bounded set $B$, $\mathrm{supp}(\psi_n) \cap B = \varnothing$.*

The proof of this may also be found in Hislop and Sigal [29]. We can now fully classify the spectra of the Hamiltonians we are interested in. Note that all electromagnetic potentials go to 0 at $x = \infty$, so all force fields are expected to approach 0 at $\infty$. In this case, we have the following:

**Theorem 5.1.8.** *Let $V : \mathbb{R}^3 \to \mathbb{R}$ be a continuous function with $V(x) \to 0$ as $|x| \to \infty$. Then:*

1. *$H = -\Delta + V$ is self-adjoint on $L^2(\mathbb{R}^{3N})$.*

2. *$\sigma_c(H) = [0, \infty)$.*

*Proof.* Verifying that $H$ is self-adjoint can be found in Hislop and Sigal [29]; the main challenge is checking that the domains match, since $H$ is manifestly symmetric.

For the second part, note that

$$||(H - \lambda)\psi_n|| - ||V\psi_n|| \leqslant (-\Delta - \lambda)\psi_n|| \leqslant ||(H - \lambda)\psi_n|| + ||V\psi_n||$$

using the triangle inequality. If the $\psi_n$ form a spreading sequence, then $||V\psi_n|| \to 0$ by the given condition, which is true iff $\lambda \in \sigma_c(-\Delta)$. We've already computed that $\sigma_c(-\Delta) = \sigma_c(H) = [0, \infty)$. $\qquad\square$

### 5.1.4 Spectra and Dynamics

Now that we have derived the spectrum of our Hamiltonian, we'll use it to understand the evolution of our system. Spectral theory tells us that $L^2(\mathbb{R}^3)$ is the direct sum of the span of eigenfunctions of $H$ Span(eigenfunc) and their orthogonal complement Span(eigenfunc)$^\perp$; further, in Span(eigenfunc)$^\perp$, $H$ has a purely continuous spectrum.

Suppose our initial wavefunction $\psi_0 \in$ Span(eigenfunc). We know that $\psi = e^{-\frac{i}{\hbar}Ht}\psi_0$. We claim that for any $\epsilon > 0$ there exists an $R$ such that $\inf_t \int_{|x|\leqslant R} |\psi|^2 \geqslant 1 - \epsilon$.

*Proof from Gustafson and Sigal [23].* Write $\psi = \sum_j a_j \psi_j$ given an orthonormal set of eigenfunctions $\psi_j$ with $H\psi_j = \lambda_j \psi_j$ and coefficients $a_j$. Assume for now that this sum is a finite sum. Then $\psi = \sum_{j=1}^{N} e^{-\frac{i}{\hbar}\lambda_j t} a_j \psi_j$ implying that for any given $R$ we have

$$\int_{|x|\geqslant R} |\psi|^2 = \sum_{j,k=1}^{N} \overline{a}_j a_k e^{-\frac{i}{\hbar}(\lambda_k - \lambda_j)t} \int_{|x|\geqslant R} \overline{\psi}_j \psi_k \leqslant \left( \sum_{j,k=1}^{N} |\overline{a}_j a_k|^2 \right)^{\frac{1}{2}} \left( \sum_{j,k=1}^{N} \left| \int_{|x|\geqslant R} \overline{\psi}_j \psi_k \right|^2 \right)^{\frac{1}{2}}$$

by the Cauchy-Schwarz inequality. The first factor is just $\int |\psi_0|^2$ by the definition of the $a_j$. For any $\epsilon > 0$, we can pick $R$ such that $\left| \int_{|x|>R} \overline{\psi}_j \psi_k \right| \leqslant \frac{1}{N} \frac{\epsilon}{\int |\psi_0|^2}$ allowing us to bound $\int_{|x| \geqslant R} |\psi|^2 \leqslant \epsilon$ as desired. If our sum is infinite, a continuity/convergence argument completes the proof. □

Thus such a $\psi$ remains essentially localized for all time and is known as a **bound state**.

On the other hand, if $\psi_0 \in \mathrm{Span}(\mathrm{eigenfunc})^{\perp}$, then for all $R$, $\int_{|x| \leqslant R} |\psi|^2 \to 0$ as $t \to \infty$ where the convergence is in the sense of the ergodic mean. This result is known as the Ruelle theorem and such a state $\psi$ is known as a **scattering state**, as it eventually leaves any fixed ball of space. A proof of the Ruelle theorem may be found in Hislop and Sigal [29].

Suppose we are now analyzing a molecular system. Positive energy systems make no physical sense, since the rest state of all electrons/protons being infinitely far away would be favorable. By Theorem 5.1.8 and the above analysis, we may safely ignore such states, i.e. all scattering states. We are thus left with looking at the bound states. Once we compute the bound states of our system, we may use a calculation similar to that of the proof above to simulate our system by projecting onto the appropriate eigenfunctions. Thus the problem of simulating our system reduces to finding eigenstates of the Hamiltonian, with eigenvalues known as eigenenergies. This leads us to the time-independent Schrödinger-Equation (TISE)

$$H\psi = E\psi. \tag{5.6}$$

The goal of the field of quantum chemistry is to solve this equation. As we will see, identifying the correct force field for MD simulation requires that one make a significant approximation to this equation.

### 5.1.5 Multi-particle Systems

The generalization to multi-particle systems has a few complications that are important to note. The Hamiltonian becomes

$$H = -\sum_{i=1}^{N} \frac{\hbar^2}{2m_i} \Delta_i + V$$

where $V : \mathbb{R}^{3N} \to \mathbb{R}$ is now the potential of all the particles. However, our state space has now changed. An important physical principle known as the Pauli exclusion principle requires that all wavefunctions of fermions (electrons, protons, and neutrons) be antisymmetric, i.e. $\psi(x_1, x_2) = -\psi(x_2, x_1)$. Thus our state space is now a subspace of $L^2(\mathbb{R}^{3N})$, not the whole space. This antisymmetrization also makes it very difficult to separate out the electrons while solving the Schrödinger Equation exactly.

The spectral analysis above also becomes more complicated in the multi-particle case [23]. It is important to separate out the motion of the center of mass from that of the rest of the system. Once that is done, results similar to Theorem 5.1.8 hold for molecular systems and in particular we always have an infinite number of bound states. The scattering states change considerably, but we will not be concerned with those for our purposes.

Thus solving Equation 5.6 is a very difficult problem, since it is a system of $3N$ coupled partial differential equations, where $N$ is the number of relevant particles, and we cannot easily separate it into simpler differential equations in general. Therefore, special methods must be used in order to understand solutions. These algorithms are the field of quantum chemistry; in general, they will only find ground-state energies accurately. Fortunately, we'll see that that is all we need for MD simulations.

## 5.2 Connecting Quantum Mechanics and Force Fields

Now that we have developed the basic quantum mechanical theory to understand how molecules actually behave, we will turn to understanding exactly how force fields and molecular dynamics approximates the true quantum-mechanical behavior of a system of molecules, which will help us understand potential sources of error in developing force fields. We will find that computing ground states of chemical systems will be important in determining force fields. The goal of this section is to discuss errors associated in connecting force fields with quantum mechanics and provide a rigorous bound on errors from nuclear quantum effects, one type of error for which mathematical bounds are already known.

For simplicity in the following sections, assume that the $N$ nuclei have the same mass $M$ and the electrons have mass $m$. The differing masses add a few additional scaling factors of $M^{\frac{1}{2}}$ but do not significantly affect the results.

### 5.2.1 The Born-Oppenheimer Approximation

This discussion largely follows Heller [27].

The fundamental idea behind the Born-Oppenheimer approximation is that nuclei are in general much more massive than electrons, i.e. $M \gg m$. Given this, consider the exact Hamiltonian

$$H(\vec{R}, \vec{r}) = K_{\text{nuc}}(\vec{R}) + K_{\text{elec}}(\vec{r}) + V(\vec{R}, \vec{r})$$

where $\vec{R}$ is the nuclear position and $\vec{r}$ is the electronic position. Explicitly, we have

$$
\begin{aligned}
K_{\text{nuc}}(\vec{R}) &= -\frac{\hbar^2}{2M}\Delta_{\vec{R}} \\
K_{\text{elec}}(\vec{r}) &= -\frac{\hbar^2}{2m}\Delta_{\vec{r}} \\
V(\vec{R}, \vec{r}) &= \sum_{i<j}\frac{k}{|\vec{R}_i - \vec{R}_j|} + \sum_{i<j}\frac{k}{|\vec{r}_i - \vec{r}_j|} - \sum_{i,j}\frac{k}{|\vec{R}_i - \vec{r}_j|}
\end{aligned}
\tag{5.7}
$$

by Coulomb's law (ignoring relativistic and spin-related corrections). As discussed previously, we now have an infinite number of eigenfunctions of $H$, since $H$ is self-adjoint. We will assume that our initial state is a bound state.

Suppose that $\Psi_k(\vec{R}, \vec{r})$ constitute all the eigenstates of the Hamiltonian $H$ with eigenenergies $E_k$. For every $\Psi_k$, we may write

$$\Psi_k(\vec{R}, \vec{r}) = \sum_{i,j} c_{kij}\psi_i(\vec{R}; \vec{r})\chi_{ij}(\vec{R}) \tag{5.8}$$

where $\psi_i(\vec{R}; \vec{r})$ is the $i$th electronic wavefunction for a given nuclear geometry $\vec{R}$ (hence a function with parametric dependence on both $i$ and $\vec{R}$) and $\chi_{ij}(\vec{R})$ is the $j$th nuclear wavefunction on the $i$th electronic potential surface determined by $\psi_i(\vec{R}; \vec{r})$. This expansion is always true; it is a consequence of completeness of the product of electronic and nuclear wavefunctions. To avoid convergence issues in the subsequent steps, it's common to approximate the sums as finite since we need to do so for numerical computations anyway. We will further assume that $\psi_i(\vec{R}; \vec{r})$ change smoothly with the nuclear coordinates $\vec{R}$.

Rearrange Equation 5.8 to get

$$\Psi_k(\vec{R}, \vec{r}) = \sum_i \psi_i(\vec{R}; \vec{r})\phi_{ki}(\vec{R}), \quad \phi_{ki}(\vec{R}) = \sum_j c_{kij}\chi_{ij}(\vec{R}). \tag{5.9}$$

$\psi_i(\vec{R}; \vec{r})$ must satisfy the TISE for the electronic Hamiltonian from Equation 5.7, so

$$(K_{\text{elec}}(\vec{r}) + V(\vec{R}, \vec{r}))\psi_i(\vec{R}; \vec{r}) = \lambda_i(\vec{R})\psi_i(\vec{R}, \vec{r}) \tag{5.10}$$

for some electronic eigenenergies $\lambda_i(\vec{R})$. We want to figure out what equation $\phi_{ki}(\vec{R})$ must satisfy.

Plugging in Equation 5.9 to the TISE and using Equation 5.10, we have

$$\sum_i (K_{\text{nuc}}(\vec{R}) + \lambda_i(\vec{R}))\psi_i(\vec{R}; \vec{r})\phi_{ki}(\vec{R}) = E_k \sum_i \psi_i(\vec{R}; \vec{r})\phi_{ki}(\vec{R}).$$

Define the electronic inner product, denoted by $\langle\rangle_r$, to be the same as the normal inner product but integrating only over $\vec{r}$, not $\vec{R}$. We then have

$$\left\langle \psi_{i'}(\vec{R}; \vec{r}) \middle| K_{\text{nuc}}(\vec{R}) \middle| \sum_i \psi_i(\vec{R}; \vec{r})\phi_{ki}(\vec{R}) \right\rangle_r + \sum_i \lambda_i(\vec{R})\phi_{ki}(\vec{R})\delta_{i'i} = \sum_i E_k\phi_{ki}(\vec{R})\delta_{i'i}$$

by normalization. Simplifying, we get

$$\left\langle \psi_{i'}(\vec{R}; \vec{r}) \middle| K_{\text{nuc}}(\vec{R}) \middle| \sum_i \psi_i(\vec{R}; \vec{r})\phi_{ki}(\vec{R}) \right\rangle_r = (E_k - \lambda_{i'}(\vec{R}))\phi_{ki'}(\vec{R}). \tag{5.11}$$

The LHS of Equation 5.11 has potentially non-diagonal terms, while the RHS is entirely diagonal with respect to the $i$ coordinates. Specifically, we may expand the LHS to

$$-\frac{\hbar^2}{2M} \left\langle \psi_{i'}(\vec{R}; \vec{r}) \middle| \sum_i (\Delta_{\vec{R}}\psi_i(\vec{R}; \vec{r}))\phi_{ki}(\vec{R}) + 2\sum_i \nabla_{\vec{R}}\psi_i(\vec{R}; \vec{r}) \cdot \nabla_{\vec{R}}\phi_{ki}(\vec{R}) + \sum_i \psi_i(\vec{R}; \vec{r})(\Delta_{\vec{R}}\phi_{ki}(\vec{R})) \right\rangle_r.$$

The **Born-Oppenheimer approximation** consists of ignoring the contributions from the first two terms of this expansion. These terms are thought to be small since nuclei, much more massive than electrons, move much more slowly than electrons and thus

electronic states change slowly with nuclear position. Applying this approximation yields the equation

$$-\frac{\hbar^2}{2m}\nabla_{\vec{R}}\phi_{ki'}(\vec{R}) = (E_k - \lambda_{i'}(\vec{R}))\phi_{ki'}(\vec{R}).$$

This is a Schrödinger equation on a modified potential surface $\lambda_{i'}(\vec{R})$. Thus we have separated electronic and nuclear motion in our wavefunction.

Assuming the Born-Oppenheimer approximation implies that we may understand the dynamics of chemical molecules by first solving for the electronic energy states $\lambda_{i'}(\vec{R})$ and then simulating nuclear motion on these modified energy states. This is very similar to MD, where we classically simulate nuclear motion on a background potential determined by electronic structure. Thus one of the most natural approaches to force field development is to compute electronic structure ground states, use that as a force field, and run classical simulations on this potential surface. This is the approach that we will examine closely in the remainder of this chapter.

We note that the Born-Oppenheimer approximation does break down when the diagonal terms are nontrivial. These occur primarily at conical intersections or potential crossings, where ground and excited state electronic energy surfaces intersect or come close to each other. These situations are uncommon chemically, but a discussion of them can be found in Heller [27] and an examination of their effect on the accuracy of MD simulations can be found in Bayer et al. [5].

### 5.2.2 Errors in Developing Born-Oppenheimer Force Fields

As outlined above, the most common approach to developing force fields for molecular dynamics simulations is by attempting to compute $\lambda_0(\vec{R})$, the ground-state electronic energy. In subsequent sections, we will examine the error resulting from this approach. However, we would like to first discuss some of the challenges surrounding performing this computation and potential sources of error that may result. A brief summary of these challenges follows; note that the exact error will depend on the force field in question [35]:

1. **Errors in quantum chemistry computations.** As noted above, it's impossible to compute $\lambda_0(\vec{R})$ exactly, thus approximate techniques using either the variational method or density functional theory must be relied on. The accuracy of these quantum chemistry computations inherently limits the accuracy of the resulting force field.

2. **Chemical transferability.** It is generally unfeasible to perform these quantum chemistry computations on large systems and undesirable to repeat them for every system under simulation. Thus scientists assume that smaller systems with similar molecules (and thus a similar chemical environment around each atom) will result in similar force fields as the larger system; this assumption is known as chemical transferability. While in general it is reasonable (electronic ground state energies rarely depend on far-away nuclei that don't affect the potential much), determining what the similar molecules should be is difficult and often a source of error.

3. **Functional forms of force fields.** The common functional forms of force fields introduced in Chapter 1 were not rigorously derived from quantum mechanics; they are approximations for empirical data that are generally not very accurate. While it's not difficult with modern computational curve-fitting to fit a curve of this form to any set of quantum chemistry data, usage of the wrong functional form can lead to large errors in the force field, especially when combined with chemical transferability. Modern force fields often use more exotic and more complicated functional forms to try to rectify this error.

4. **Multi-body effects.** In particular, multi-body effects, or effects caused by interactions of multiple molecules that are not accounted by pairwise additive potentials, have proven difficult to model at any stage. Quantum chemistry calculations do not model them well and most force fields ignore them entirely due to the computational cost.

A variety of methods have been developed to circumvent some of these issues by combining quantum mechanics and molecular dynamics in various ways. QM/MM methods attempt to model certain parts of the system quantum mechanically and the rest using molecular dynamics, and purely *ab initio* molecular dynamics explicitly runs electronic structure computations at each stage of the simulation to compute $\lambda_0(\vec{R})$ and determine the appropriate force without using a force field. These methods are significantly more computationally expensive than MD, and thus much less popular.

It's also important to note that some more modern force field developers do not attempt to determine $\lambda_0(\vec{R})$ at all, but use experimental parameters to determine constants in certain functional forms. It is extremely difficult to make rigorous statements about the accuracy of these force fields.

We will now consider errors that result even from using a completely accurate $\lambda_0(\vec{R})$ to measure an observable. All methods derived from quantum mechanics will fall prey to these errors.

### 5.2.3 Wigner Transforms and Weyl Quantization

Suppose we could accurately determine $\lambda_0(\vec{R})$. Suppose also that the Born-Oppenheimer approximation holds reasonably well (i.e. we are far from a conical intersection or something similar). The remaining errors in simulating our system according to classical molecular dynamics as opposed to the Schrödinger equation result from the fact that the nuclei themselves are also quantum-mechanical in nature and do not exactly follow the classical equations of motion. These errors are thus considered **nuclear quantum effects**. For the remainder of this thesis, we will explore some mathematically proven bounds on the size of nuclear quantum effects and then my current research into ways to account for these effects. We will first introduce some auxiliary useful quantum mechanics.

For convenience and for the next two sections only we will now change to units where $\hbar = 1$. We will also add some extra factors of $M^{-\frac{1}{2}}$ to account for scaling the coordinates to make the masses identical, as noted above.

For each observable $A$, we need to define an associated quantum operator $\hat{A}$ to use to compute the quantum expectation. There are many possible choices, known as quantiza-

tions, and it can be verified that these are all equivalent in the classical limit. We will use **Weyl quantization**. Given a wavefunction $\phi(\vec{R})$, we have

$$\hat{A}\phi(\vec{R}) = \left(2\pi M^{-\frac{1}{2}}\right)^{-3N} \int e^{iM^{\frac{1}{2}}(\vec{R}-\vec{R}')\cdot\vec{p}} A\left(\frac{\vec{R}+\vec{R}'}{2}, \vec{p}\right)\phi(\vec{R}')\,d\vec{R}'\,d\vec{p}. \qquad (5.12)$$

One can check that this does reduce to the classical limit as desired. Our expectation value for our observable $A$ is now $\langle A \rangle = \langle \phi(\vec{R})|\hat{A}\phi(\vec{R})\rangle$ as before.

We choose this quantization for the following important reason. We may define the **Wigner function**

$$W(\vec{R}, \vec{p}) = \left(2\pi M^{-\frac{1}{2}}\right)^{-3N} \int e^{iM^{\frac{1}{2}}\vec{R}'\cdot\vec{p}} \left\langle \phi\left(\vec{R}+\frac{\vec{R}'}{2}\right) \middle| \phi\left(\vec{R}-\frac{\vec{R}'}{2}\right)\right\rangle d\vec{R}'. \qquad (5.13)$$

This function has a number of important properties. It can be verified that integrating out momentum yields the position probability distribution $\int W(\vec{R}, \vec{p})\,d\vec{p} = |\phi(\vec{R})|^2$ and similarly for integrating out position. Thus the Wigner function represents something similar to a joint probability distribution in phase space. It isn't quite a probability distribution, since it can be negative; thus it is occasionally called a quasi-probability distribution [10].

Most usefully for us, it can be shown that

$$\langle A \rangle = \langle \phi(\vec{R})|\hat{A}\phi(\vec{R})\rangle = \int A(\vec{R}, \vec{p})W(\vec{R}, \vec{p})\,d\vec{R}\,d\vec{p} \qquad (5.14)$$

by grinding through the relevant integrals. Thus we can easily measure the expectation value of our observable defined via Weyl quantization by using the Wigner function as a probability distribution. This is particularly convenient since it will make it easy for us to connect the quantum value of the observable to its classically measured value.

There's another less significant issue we need to fix prior to stating our theorem. We are not guaranteed that $\lambda_0(\vec{R})$ is a smooth function of $\vec{R}$. To fix this, define

$$\lambda_\delta(\vec{R}) = (\pi\delta^2)^{\frac{3N}{2}} \int \lambda_0(\vec{R}-\vec{R}')e^{-\frac{|\vec{R}'|^2}{\delta^2}}\,d\vec{R}'$$

which is now clearly a smooth function that approximates $\lambda_0$ with error $\mathcal{O}(\delta)$, assuming $\lambda_0$ is Lipschitz continuous, i.e. $\lambda_0(\vec{R}) - \lambda_0(\vec{R}') \leqslant C|\vec{R}-\vec{R}'|$. This is because

$$\lambda_\delta(\vec{R}) - \lambda_0(\vec{R}) = (\pi\delta^2)^{\frac{3N}{2}} \int \left(\lambda_0(\vec{R}-\vec{R}') - \lambda_0(\vec{R})\right)e^{-\frac{|\vec{R}'|^2}{\delta^2}}\,d\vec{R}' \leqslant C(\pi\delta)^{-\frac{3N}{2}}\int |\vec{R}'|e^{-\frac{|\vec{R}'|^2}{\delta^2}}\,d\vec{R}' = \mathcal{O}(\delta).$$

We'll work on $\lambda_\delta$ instead.

Given an observable $A$ and a flow computed via molecular dynamics $S_t(\vec{R}_0, \vec{p}_0)$, we may define

$$A_{\text{MD}}(\vec{R}_0, \vec{p}_0) = \lim_{T\to\infty} \frac{1}{T}\int_0^T A(S_t(\vec{R}_0, \vec{p}_0))\,dt$$

as the value computed by an accurate MD simulation. We ignore time discretization error here and assume that sampling error converges as $\mathcal{O}(T^{-\gamma})$ for some $\gamma > 0$. We'd like to compare this value to

$$A_\delta = \int A(\vec{R}, \vec{p})W(\vec{R}, \vec{p})\,d\vec{R}\,d\vec{p}$$

as in Equation 5.14.

### 5.2.4   Bounding Error from Nuclear Quantum Effects

We will now state the main result about the error in nuclear quantum effects.

**Theorem 5.2.1.** *Under the conditions above, and assuming that $A$, $A_{\mathrm{MD}}$, and $V$ are smooth and grow with at most polynomial growth for large $\vec{R}$, $\vec{p}$, and that the eigenfunctions decay faster than polynomials in $\vec{R}$ outside a compact domain, we have the weighted error estimate*

$$\int (A_\delta - A_{\mathrm{MD}}(\vec{R}_0, \vec{p}_0)) W_\delta(\vec{R}_0, \vec{p}_0)\, d\vec{R}_0\, d\vec{p}_0 = \mathcal{O}(T(M^{-1} + \delta^{-1}M^{-2}) + T^{1-\gamma}M^{-\frac{1}{2}} + T^{-\gamma}).$$
(5.15)

*Proof Outline from Bayer et al. [5].* We omit some of the messy computations in what follows; details may be found in Bayer et al. [5].

The TDSE, as noted before, is solved by the operator $e^{-iM^{\frac{1}{2}}t\hat{H}_\delta}$. This must therefore fix observables associated with the eigenfunction $\Phi_\delta$. So we have

$$A_\delta = \int \overline{\Phi}_\delta(\vec{R})\hat{A}\Phi_\delta(\vec{R})\, d\vec{R} = \int \Phi_\delta(\vec{R})e^{iM^{\frac{1}{2}}t\hat{H}_\delta}\hat{A}e^{-iM^{\frac{1}{2}}t\hat{H}_\delta}\Phi_\delta(\vec{R})\, d\vec{R}$$

using Equation 5.12. We also have

$$\int A_{\mathrm{MD}}(\vec{R}_0, \vec{p}_0)W(\vec{R}_0, \vec{p}_0)\, d\vec{R}_0\, d\vec{p}_0 = \lim_{T\to\infty} \frac{1}{T} \int_0^T \int A(S_t(\vec{R}_0, \vec{p}_0))W_\delta(\vec{R}_0, \vec{p}_0)\, d\vec{R}_0\, d\vec{p}_0\, dt$$

$$= \lim_{T\to\infty} \frac{1}{T} \int_0^T \int \Phi_\delta(\vec{R}_0)\widehat{A \circ S_t}\Phi_\delta(\vec{R}_0)\, d\vec{R}_0\, dt$$

by Equation 5.14.

To compare these, note that $S_0$ is the identity. We then write their difference as

$$\int \Phi_\delta(\vec{R}_0) \left[ \int_0^t \frac{d}{ds} \left( e^{iM^{\frac{1}{2}}(t-s)\hat{H}_\delta} \widehat{A \circ S_s} e^{-iM^{\frac{1}{2}}(t-s)\hat{H}_s} \right) ds \right] \Phi_\delta(\vec{R}_0)\, d\vec{R}_0.$$
(5.16)

Using the fact that $\frac{d}{ds} A \circ S_s = \frac{d}{dr} A(S_s(\vec{R}_r, \vec{p}_r))|_{r=0}$, we may compute

$$\frac{d}{ds} \left( e^{iM^{\frac{1}{2}}(t-s)\hat{H}_\delta} \widehat{A \circ S_s} e^{-iM^{\frac{1}{2}}(t-s)\hat{H}_s} \right)$$

$$= e^{iM^{\frac{1}{2}}(t-s)\hat{H}_\delta} \left( \mathrm{Op}[((\vec{p}\cdot\vec{\nabla}_{\vec{R}} - \vec{\nabla}\lambda_\delta(\vec{R})\cdot\vec{\nabla}_{\vec{p}})A) \circ S_s] - iM^{\frac{1}{2}}[\hat{H}_\delta, \widehat{A \circ S_s}] \right) e^{-iM^{\frac{1}{2}}(t-s)\hat{H}_\delta}$$
(5.17)

where the square brackets are the commutator and Op means the Weyl quantization.

Plugging in Equation 5.17 into Equation 5.16, we can clean things up by noting that $\Phi_\delta(\vec{R}_0)$ is an eigenfunction of $\hat{H}_\delta$ so the outer exponentials must cancel out. So we focus on the operator in the middle. Applying the Moyal expansion to the commutator of two Weyl operators (see details in Bayer et al. [5]), we find that

$$\mathrm{Op}[((\vec{p}\cdot\vec{\nabla}_{\vec{R}} - \vec{\nabla}\lambda_\delta(\vec{R})\cdot\vec{\nabla}_{\vec{p}})A) \circ S_s] - iM^{\frac{1}{2}}[\hat{H}_\delta, \widehat{A \circ S_s}] = \hat{R}_M(s)$$
(5.18)

where the remainder term $\hat{R}_M$ is

$$\hat{R}_M(s) = \mathrm{Op}\left[\sum_{n=1}^{\infty} 2M^{-n}(2i)^{-n} \sum_{|\alpha|=n} \frac{(-1)^{|\alpha|}}{\alpha!} \partial_{\vec{R}_0}^{\alpha} \lambda_\delta(\vec{R}_0) \partial_{\vec{p}_0}^{\alpha} (A \circ S_s)(\vec{R}_0, \vec{p}_0)\right]. \qquad (5.19)$$

Combining Equations 5.18 and 5.19 shows that the error

$$\int \left(A_\delta - \frac{1}{T}\int_0^T A \circ S_t(\vec{R}_0, \vec{p}_0)\, dt\right) W_\delta(\vec{R}_0, \vec{p}_0)\, d\vec{R}_0\, d\vec{p}_0 = T^{-1}\int_0^T \int_0^t \int \hat{R}_M(s) W_\delta(\vec{R}_0, \vec{p}_0)\, d\vec{R}_0\, d\vec{p}_0\, ds\, dt.$$

$$(5.20)$$

This is almost the error we want to compute, up to the previously noted bound of $\mathcal{O}(T^{-\delta})$ in the convergence of sampling error. Thus we'll focus on bounding this error, and in particular on bounding $\hat{R}_M$.

To bound $\hat{R}_M$, substitute $A \circ S_s \mapsto A_{\mathrm{MD}} + (A \circ S_s - A_{\mathrm{MD}})$. Since $\lambda_0$ is Lipschitz continuous and $A_{\mathrm{MD}}$ grows at most polynomial in $\vec{R}$, we may bound the terms in Equation 5.20. The $n = 2$ term is bounded by $\mathcal{O}(\delta^{-1}M^{-2})$, and all higher-order terms are bounded by $\mathcal{O}(\delta^{1-n}M^{-n})$, which is suppressed by the above bound; the $n = 1$ term is bounded by $\mathcal{O}(M^{-1})$. Thus we integrate to get a bound of $\mathcal{O}(TM^{-1} + TM^{-2}\delta^{-1})$. The term with $A \circ S_s - A_{\mathrm{MD}}$ is harder since $A \circ S_s$ may grow exponentially. But the difference is bounded by $\mathcal{O}(T^{-\gamma})$ as we proved above. Integrate by parts in $\vec{p}_0$ to move derivatives to the Wigner function. This gives us a bound of $\mathcal{O}(T^{1-\gamma}M^{-\frac{1}{2}})$ (see details in Bayer et al. [5]). Combining these results gives us the desired bound. $\qquad \square$

We have thus proven that there is an error term associated with nuclear quantum effects, even with Born-Oppenheimer force fields, but that it can be bounded. However, this bound does not guarantee convergence to 0 at large $T$, especially for small masses $M$ like the mass of the hydrogen nucleus. It seems conceivable that there might be another way of computing a force field that does not include such errors. We now turn to the subject of my own research and to the last chapter of this thesis: force-field functor theory, which will provide a way for us to answer this question.

# Chapter 6

# Force-Field Functor Theory

In the last chapter, we asserted that the correct force field should look something like the lowest-energy electronic eigenstate $\lambda_0(\vec{R})$ and proved error bounds based on this. But while it is true that physical intuition indicates this is what the force field should look like and that in the limit of infinite nuclear mass this is accurate, we still see a significant error term associated with nuclear quantum effects for finite nuclear mass. There have been many computational methods developed to reduce this error [11, 42, 45, 40, 48, 34, 30, 19, 47, 12, 20, 24]. But these methods all fundamentally change the MD algorithm itself, usually at significant computational cost.

We propose an alternative approach: develop a different force field that accurately accounts for nuclear quantum effects. In **force-field functor theory**, we map a Born-Oppenheimer force field $V(\vec{R}) = \lambda_0(\vec{R})$ to an effective force field $W(\vec{R})$ that will account for nuclear quantum effects; this map $\mathcal{F} : V \to W$ is known as the force-field functor. In this chapter, we will prove the uniqueness and existence of this map and then discuss an efficient way to compute it given a Born-Oppenheimer force field (since that is the computable quantity in quantum chemistry).

Force-field functor theory applies at a constant temperature, not a constant energy. Thus our simulations will now be NVT simulations, not NVE simulations. Further, not all of our particles will be in the same state corresponding to the same wavefunction; they will instead be in a mixture of states to maintain a constant temperature. Thus we will start by briefly introducing some of the relevant physics behind modeling constant-temperature quantum mechanical systems and then proceed to force-field functor theory.

We will reintroduce factors of $\hbar$ for this chapter as they prove to be important and let the masses of our particles possibly be different.

## 6.1  Density Matrices and Force-Field Functor Existence

### 6.1.1  The Density Matrix

Return to the one-particle case for now. Suppose we have a Hamiltonian $H$ on the Hilbert space $L^2(\mathbb{R}^3)$. Pick an orthonormal basis $\psi_j$. Then any state $\phi$ can be written as $\phi = \sum_j a_j \psi_j$ and any observable $A$ can be computed as $\langle A \rangle = \langle \psi | A \psi \rangle = \sum_{m,n} \bar{a}_m a_n \langle \psi_m | A \psi_n \rangle$.

What if we only know the system is in state $\psi_n$ with probability $p_n$? Then our average is given by the expression actually looks like $\langle A \rangle = \sum_n p_n \langle \psi_n | A \psi_n \rangle$. Define the **density matrix** operator

$$\rho = \sum_n p_n |\psi_n\rangle\langle\psi_n| \tag{6.1}$$

so we then have

$$\langle A \rangle = \text{Tr}(A\rho). \tag{6.2}$$

Note that $\rho$ is a positive trace class operator with $\text{Tr}\,\rho = 1$ by normalization. Any such operator (not just one of the form above) corresponds to a generalized state and is known as a density matrix [23].

Using the Schrödinger Equation, it's easy to show that the equation of state for $\rho$ is

$$i\frac{\partial\rho}{\partial t} = \frac{1}{\hbar}[H,\rho],\tag{6.3}$$

known as the Landau-von Neumann equation.

In our case, $H$ has point spectrum, so any operator $f(H)$ for a positive function $f$ is a density matrix. Thus $\rho = f(H)$ is a time-independent solution of Equation 6.3 since $[f(H),H] = 0$. Thus positive functions of the Hamiltonian will be stationary solutions.

States that are of the form $\rho = |\psi\rangle\langle\psi|$ correspond directly to wavefunctions $\psi$ and are therefore known as pure states, i.e. the probability for the system to be in the state $\psi$ is 1. The system can also be in a mixed state, where the population of each pure state is given by a probability distribution, representing the fact that we have incomplete information about our system and do not know that it is in any one pure state [23].

It is also possible to take the Wigner transform of a density matrix. To do this, we define the Weyl transform, the inverse of Weyl quantization. For an operator $\hat{A}$, we have

$$\tilde{A}(\vec{x},\vec{p}) = \int e^{-\frac{i}{\hbar}\vec{p}\cdot\vec{y}}\left\langle \vec{x}+\frac{\vec{y}}{2}\,\middle|\,\hat{A}\,\middle|\,\vec{x}-\frac{\vec{y}}{2}\right\rangle d\vec{y}.\tag{6.4}$$

Now, we define

$$W(\vec{x},\vec{p}) = \frac{\tilde{\rho}}{2\pi\hbar}.\tag{6.5}$$

One can verify using Equation 6.4 that this is the same Wigner transform as we previously defined

$$W(\vec{x},\vec{p}) = \frac{1}{2\pi\hbar}\int e^{-\frac{i}{\hbar}\vec{p}\cdot\vec{y}}\psi\left(\vec{x}+\frac{\vec{y}}{2}\right)\overline{\psi}\left(\vec{x}-\frac{\vec{y}}{2}\right)d\vec{y}.$$

## 6.1.2 Thermodynamics and the Density Matrix

The density matrix is useful for understanding thermodynamic systems, or systems with large numbers of molecules in which we do not know exact information about all the initial positions of the molecules. Suppose we have $N$ particles that interact pairwise in a volume $\Lambda$. We have a Hamiltonian $H_{N,\Lambda}$, and we would like to take the conceptual limit $N \to \infty$. Clearly as noted above $f(H_{N,\Lambda})$ for any positive $f$ that decays sufficiently fast is a stationary state.

But which stationary states are stable? Most stationary states are not asymptotically stable, so they do not provide useful physical information about our system. However, some states are asymptotically stable and convergence to such states is very rapid. To understand these states, we will work in the conceptual limit of $N = \infty, \Lambda = \infty$ but fix $\frac{N}{|\Lambda|}$. This is the **thermodynamic limit**, and the study of such states is the study of thermodynamics.

In the absence of an external environment, stationary states with infinitely many particles are called **thermal equilibrium states**. They are computed by the second law of

thermodynamics, which says that $\rho$ is an equilibrium state iff it maximizes the **von Neumann entropy** $S(\rho)$ while fixing the energy $E(\rho) = E$. The entropy is

$$S(\rho) = -\operatorname{Tr}(\rho \log \rho) \tag{6.6}$$

where log is defined via power series. The energy is as expected $E(\rho) = \operatorname{Tr}(H\rho)$.

Optimizing the entropy is a Lagrange multiplier problem in variational calculus. We may instead minimize the **Helmholtz free energy**

$$F(\rho) = E(\rho) - TS(\rho) \tag{6.7}$$

for a Lagrange parameter $T$ called the **temperature**. We will also occasionally see the inverse $\beta = \frac{1}{k_B T}$ for a constant $k_B$ known as **Boltzmann's constant**. Minimizing this yields the family

$$\rho_T = \frac{e^{-\beta H}}{Z(T)}, Z(T) = \operatorname{Tr}[e^{-\beta H}]. \tag{6.8}$$

The function $Z(T)$ is known as the **partition function** and can be used to derive many important properties of the system [23]. For example,

$$F(T) = F(\rho_T) = -T \log Z(T).$$

Thus we now understand what equilibrium states we expect to see quantum-mechanically at a given temperature $T$. It's important to note that the coefficients $e^{-\beta E_n}$ for each eigenenergy $E_n$ associated with an eigenstate $\psi_n$ represent the probability of finding a particular particle in our infinite-particle system in the one-particle state associated with $\psi_n$. There is then a second layer of quantum-mechanical probability associated with determining the value of any observable given the particle in the state $\psi_n$.

In the classical limit, the formalism above largely still applies. However, our states are now particular position/momentum points, i.e. points in phase space, as opposed to wavefunctions.

### 6.1.3 Existence and Uniqueness of the Force-Field Functor

We now present the main foundational result of force-field functor theory, which guarantees the existence and uniqueness of a classical potential that reproduces the quantum equilibrium position-space probability distribution. This is the best we can possibly do with an MD simulation, since the momentum-space probability distribution is fixed from the beginning as the force field is momentum-independent.

**Theorem 6.1.1.** *Given a Born-Oppenheimer force field $V(\vec{x})$ and a temperature $T$, there exists a unique effective force field $W(\vec{x}) = \mathcal{F}[V(\vec{x})]$ such that the classical position-space probability distribution $\eta_c(\vec{x})$ computed from $W(\vec{x})$ is equal to the quantum position-space probability distribution $\eta_q(\vec{x})$ computed from $V(\vec{x})$.*

*Proof from Babbush et al. [3].* Existence is relatively simple. We may compute our equilibrium density $\rho_T$ as above, and then the quantum probability position-space distribution is

$$\eta_q(\vec{x}) = \langle \vec{x} | \rho_T \vec{x} \rangle \tag{6.9}$$

and the classical probability distribution of any classical potential $W(\vec{x})$ is

$$\eta_c(\vec{x}) = \frac{1}{Z}e^{-\beta W(\vec{x})}. \tag{6.10}$$

Simultaneously solving Equation 6.10 and 6.9 yields

$$W(\vec{x}) = -\frac{1}{\beta}\log\left(Z\langle\vec{x}|\rho_T\vec{x}\rangle\right) \tag{6.11}$$

which shows that the functor exists.

Proving uniqueness is a bit more difficult. We proceed in two steps: we will first show that the quantum position-space distribution is uniquely given from $V(\vec{x})$ and then that the classical potential is uniquely given from any classical position-space distribution. We begin with a lemma.

**Lemma 6.1.2** (Quantum Bogoliubov Inequality). *For positive definite $\rho$ with unit trace, we have*

$$F(\rho) > F(\rho_0)$$

*if $F$ is the Helmholtz free energy and $\rho \neq \rho_0$, where $\rho_0$ is the equilibrium distribution.*

*Proof from Babbush et al. [3].* Define

$$\rho_\lambda = \frac{e^{-\beta(H+\lambda\Delta)}}{\mathrm{Tr}\left(e^{-\beta(H+\lambda\Delta)}\right)}, \Delta = -\frac{1}{\beta}\log\rho - H.$$

Clearly $\rho_\lambda = \rho_0$ if $\lambda = 0$ and $\rho_\lambda = \rho$ if $\lambda = 1$, so $F(\rho) - F(\rho_0) = \int_0^1 \frac{\partial}{\partial\lambda}F(\rho_\lambda)\,d\lambda$.

We have

$$F(\rho_\lambda) = \mathrm{Tr}\left[\rho_\lambda\left(H + \lambda\Delta + \frac{1}{\beta}\log\rho_\lambda\right)\right] - \lambda\,\mathrm{Tr}[\Delta\rho_\lambda]$$

by definition, and the first trace is stationary since $\rho_\lambda$ is the equilibrium density matrix for $H + \lambda\Delta$. Thus we differentiate the second trace $\frac{\partial}{\partial\lambda}F[\rho_\lambda] = -\lambda\,\mathrm{Tr}\left[\Delta\frac{\partial}{\partial\lambda}\rho_\lambda\right]$.

Define $\langle X\rangle_\lambda = \mathrm{Tr}[\rho_\lambda X]$. Then we may compute

$$\frac{\partial}{\partial\lambda}\rho_\lambda = -\int_0^\beta \rho_\lambda[\Delta_\lambda(\beta') - \langle\Delta\rangle_{\lambda'}], \Delta_\lambda(\beta') = e^{\beta'(H+\lambda\Delta)}\Delta e^{-\beta'(H+\lambda\Delta)}.$$

Thus

$$\frac{\partial}{\partial\lambda}F[\rho_\lambda] = \lambda\int_0^\beta(\langle\Delta\Delta_\lambda(\beta')\rangle_\lambda - \langle\Delta\rangle_\lambda^2) = \lambda\int_0^\beta d\beta'\left\langle\left(\Delta_\lambda\left(\frac{1}{2}\beta'\right) - \langle\Delta\rangle_\lambda\right)^\dagger\left(\Delta_\lambda\left(\frac{1}{2}\beta'\right) - \langle\Delta\rangle_\lambda\right)\right\rangle_\lambda$$

using some trace identities. This integral is non-negative and can be zero iff $\Delta$ is the unit operator, i.e. if $\rho = \rho_0$. Thus we get the desired. $\square$

Now suppose that there were two potentials $\widetilde{V}(\vec{x}), V(\vec{x})$ that had the same quantum position-space probability distribution $\eta_q(\vec{x})$. We have corresponding Hamiltonians, density matrices, and free energies $\widetilde{H}, \widetilde{\rho}_0, \widetilde{F}$. Then we apply Lemma 6.1.2 to get

$$\widetilde{F} = \mathrm{Tr}\left[\widetilde{\rho}_0\left(\widetilde{H} + \frac{1}{\beta}\log\widetilde{\rho}_0\right)\right] < \mathrm{Tr}\left[\rho_0\left(\widetilde{H} + \frac{1}{\beta}\log\rho_0\right)\right] = F + \mathrm{Tr}[\rho_0\widetilde{V}(\vec{x}) - \rho_0 V(\vec{x})].$$

This last term is just $\int d\vec{x}(\widetilde{V}(\vec{x}) - V(\vec{x}))\eta_q(\vec{x})$ so we have

$$\widetilde{F} < F + \int d\vec{x}(\widetilde{V}(\vec{x}) - V(\vec{x}))\eta_q(\vec{x}).$$

Interchanging variables leads to a contradiction.

For the other step, we begin with another, similar lemma.

**Lemma 6.1.3** (Classical Bogoliubov Inequality)**.** *If $\eta_c(\vec{x})$ is the equilibrium density for a classical system and $\eta(\vec{x})$ is another different density, then*

$$F(\eta(\vec{x})) > F(\eta_c(\vec{x})).$$

*Proof from Babbush et al. [3].* Note that

$$\frac{1}{\beta}\int(\eta_c(\vec{x})\log\eta_c(\vec{x}) - \eta_c(\vec{x})\log\eta(\vec{x}))\,d\vec{x} \geqslant \frac{1}{\beta}\int(\eta_c(\vec{x}) - \eta(\vec{x}))\,d\vec{x} = 0$$

by normalization and the fact that $\log x \geqslant 1 - \frac{1}{x}$. Thus $\left\langle\frac{1}{\beta}\log\eta_c(\vec{x})\right\rangle \geqslant \left\langle\frac{1}{\beta}\log\eta(\vec{x})\right\rangle$. Plugging in the standard formulas for the number density, this yields

$$\left\langle -\widetilde{E}(\vec{x}) - \frac{1}{\beta}\log\widetilde{Z}\right\rangle \geqslant \left\langle -E(\vec{x}) - \frac{1}{\beta}\log Z\right\rangle$$

which is equivalent to the desired. $\qquad\square$

Now suppose that there were two effective potentials $\widetilde{W}(\vec{x}), W(\vec{x})$ that led to the same position-space probability distribution. Then

$$\widetilde{F} = \int\eta_c(\vec{x})\widetilde{W}(\vec{x})\,d\vec{x} + \frac{1}{\beta}\int\eta_c(\vec{x})\log\eta_c(\vec{x})\,d\vec{x} < A + \int\eta_c(\vec{x})(\widetilde{W}(\vec{x}) - W(\vec{x}))$$

applying Lemma 6.1.3. We can now complete the proof in a similar fashion to the first part. $\qquad\square$

The force-field functor theorem guarantees the existence of a unique potential $W(\vec{x})$ that classically reproduces the quantum position-space probability distribution. Since it classically reproduces the correct quantum distribution, it will also correctly reproduce all thermodynamic properties which depend solely or mostly on position. Dynamic properties that depend on momentum will not be correctly reproduced. Regardless, this is a

powerful method of guaranteeing accurate computation of observables via MD simulation.

Of course, this proof is essentially non-constructive, since there are few easy ways to determine the density matrix. We will now need to find a reasonable way of computationally approximating the effective potential given an original Born-Oppenheimer potential to work with.

Note that the initial force field must approximate or be fitted to $\lambda_0(\vec{x})$, the ground-state electronic energy term, for force-field functor theory to be valid. Thus our force field must be fit to electronic structure data. Force fields fit to experimental parameter data may already account for nuclear quantum effects in some empirical and approximate way; adding force-field functor corrections will result in double-counting of nuclear quantum effects [50].

## 6.2 The Wigner Expansion and Force-Field Functors

The proof of Theorem 6.1.1 makes clear that the quantum position-space probability distribution $\eta_q(\vec{x})$ is the key object to compute here. Computing the density matrix is very difficult, so integrating that to find $\eta_q(\vec{x})$ is not generally tractable (though this is the inspiration for many alternative approaches to nuclear quantum effects). We instead focus on the Wigner transform, which by Equation 6.5 is equivalent. Then, Theorem 6.1.1 tells us that

$$\widetilde{V}(\vec{x}) = -\frac{1}{\beta} \log \left[ \int W(\vec{x}, \vec{p}) \, d\vec{p} \right]. \tag{6.12}$$

where $\widetilde{V}(\vec{x}) = \mathcal{F}[V(\vec{x})]$ is the effective potential. In this section, we find a simple, computationally effective way of approximating the Wigner transform. We assume our potential is analytic, as it usually is for MD simulations.

Most of the following section is copied from my recent paper, Sundar et al. [51].

### 6.2.1 The Wigner Expansion

We first prove the following lemma.

**Lemma 6.2.1.** *The Wigner function $W(\vec{x}, \vec{p})$ satisfies the following equation:*

$$\frac{\partial W}{\partial t} = -\sum_k \frac{p_k}{m_k} \frac{\partial W}{\partial x_k} + \sum_{(\lambda_1, \ldots, \lambda_n)} \frac{\partial^{\lambda_1 + \ldots + \lambda_n} V}{\partial x_1^{\lambda_1} \ldots \partial x_n^{\lambda_n}} \frac{\left(\frac{\hbar}{i}\right)^{\lambda_1 + \ldots + \lambda_n - 1}}{\lambda_1! \ldots \lambda_n!} \frac{\partial^{\lambda_1 + \ldots + \lambda_n} W}{\partial p_1^{\lambda_1} \ldots \partial p_n^{\lambda_n}}. \tag{6.13}$$

*Here, $x_1, \ldots, x_n$ represent the n position coordinates and $p_1, \ldots, p_n$ represent the n momentum coordinates. $m_i$ is the associated mass for the ith position or momentum coordinate.*

*Proof from Wigner [52].* The basic approach here is to compute the derivative directly using the definition of the Wigner function and then apply the time-dependent Schrödinger

Equation and integrate by parts a few times.

$$
\begin{aligned}
\frac{\partial W}{\partial t} &= \frac{1}{2\pi\hbar} \int e^{-\frac{i}{\hbar}\vec{p}\cdot\vec{y}} \left[ \frac{\partial \psi}{\partial t}\left(\vec{x}+\frac{\vec{y}}{2}\right) \overline{\psi}\left(\vec{x}-\frac{\vec{y}}{2}\right) + \psi\left(\vec{x}+\frac{\vec{y}}{2}\right) \frac{\partial \overline{\psi}}{\partial t}\left(\vec{x}-\frac{\vec{y}}{2}\right) \right] d\vec{y} \\
&= \frac{1}{2\pi\hbar} \int e^{-\frac{i}{\hbar}\vec{p}\cdot\vec{y}} \left[ -\sum_k \frac{i\hbar}{2m_k}\left( \frac{\partial^2 \psi}{\partial x_k^2}\left(\vec{x}+\frac{\vec{y}}{2}\right)\overline{\psi}\left(\vec{x}-\frac{\vec{y}}{2}\right) + \psi\left(\vec{x}+\frac{\vec{y}}{2}\right)\frac{\partial^2 \overline{\psi}}{\partial x_k^2}\left(\vec{x}-\frac{\vec{y}}{2}\right) \right) \right.\\
&\quad \left. + \frac{i}{\hbar}\psi\left(\vec{x}+\frac{\vec{y}}{2}\right)\overline{\psi}\left(\vec{x}-\frac{\vec{y}}{2}\right)\left(V\left(\vec{x}+\frac{\vec{y}}{2}\right)+V\left(\vec{x}-\frac{\vec{y}}{2}\right)\right) \right] \\
&= \frac{1}{2\pi\hbar} \int e^{-\frac{i}{\hbar}\vec{p}\cdot\vec{y}} \left[ -\sum_k \frac{i\hbar}{2m_k}\left( 4\frac{\partial^2 \psi}{\partial y_k^2}\left(\vec{x}+\frac{\vec{y}}{2}\right)\overline{\psi}\left(\vec{x}-\frac{\vec{y}}{2}\right) - 4\psi\left(\vec{x}+\frac{\vec{y}}{2}\right)\frac{\partial^2 \overline{\psi}}{\partial y_k^2}\left(\vec{x}-\frac{\vec{y}}{2}\right) \right) \right.\\
&\quad \left. + \frac{i}{\hbar}\psi\left(\vec{x}+\frac{\vec{y}}{2}\right)\overline{\psi}\left(\vec{x}-\frac{\vec{y}}{2}\right)\left( \sum_{(\lambda_1,\dots,\lambda_n)} \frac{\partial^{\lambda_1+\dots+\lambda_n}V}{\partial x_1^{\lambda_1}\dots\partial x_n^{\lambda_n}} \frac{y_1^{\lambda_1}\dots y_n^{\lambda_n}}{\lambda_1!\dots\lambda_n!} \right) \right] \\
&= \frac{1}{2\pi\hbar} \int e^{-\frac{i}{\hbar}\vec{p}\cdot\vec{y}} \left[ -\sum_k \frac{p_k}{2m_k}\left( 4\frac{\partial \psi}{\partial y_k}\left(\vec{x}+\frac{\vec{y}}{2}\right)\overline{\psi}\left(\vec{x}-\frac{\vec{y}}{2}\right) - 4\psi\left(\vec{x}+\frac{\vec{y}}{2}\right)\frac{\partial \overline{\psi}}{\partial y_k}\left(\vec{x}-\frac{\vec{y}}{2}\right) \right) \right.\\
&\quad \left. + \psi\left(\vec{x}+\frac{\vec{y}}{2}\right)\overline{\psi}\left(\vec{x}-\frac{\vec{y}}{2}\right)\left( \sum_{(\lambda_1,\dots,\lambda_n)} \frac{\partial^{\lambda_1+\dots+\lambda_n}V}{\partial x_1^{\lambda_1}\dots\partial x_n^{\lambda_n}} \frac{\left(\frac{\hbar}{i}\right)^{\lambda_1+\dots+\lambda_n-1}}{\lambda_1!\dots\lambda_n!}\left(\frac{i}{\hbar}y_1\right)^{\lambda_1}\dots\left(\frac{i}{\hbar}y_n\right)^{\lambda_n} \right) \right] \\
&= \frac{1}{2\pi\hbar} \int e^{-\frac{i}{\hbar}\vec{p}\cdot\vec{y}} \left[ -\sum_k \frac{p_k}{2m_k}\left( 2\frac{\partial \psi}{\partial x_k}\left(\vec{x}+\frac{\vec{y}}{2}\right)\overline{\psi}\left(\vec{x}-\frac{\vec{y}}{2}\right) + 2\psi\left(\vec{x}+\frac{\vec{y}}{2}\right)\frac{\partial \overline{\psi}}{\partial x_k}\left(\vec{x}-\frac{\vec{y}}{2}\right) \right) \right.\\
&\quad \left. + \psi\left(\vec{x}+\frac{\vec{y}}{2}\right)\overline{\psi}\left(\vec{x}-\frac{\vec{y}}{2}\right)\left( \sum_{(\lambda_1,\dots,\lambda_n)} \frac{\partial^{\lambda_1+\dots+\lambda_n}V}{\partial x_1^{\lambda_1}\dots\partial x_n^{\lambda_n}} \frac{\left(\frac{\hbar}{i}\right)^{\lambda_1+\dots+\lambda_n-1}}{\lambda_1!\dots\lambda_n!}\left(\frac{i}{\hbar}y_1\right)^{\lambda_1}\dots\left(\frac{i}{\hbar}y_n\right)^{\lambda_n} \right) \right] \\
&= -\sum_k \frac{p_k}{m_k}\frac{\partial W}{\partial x_k} + \sum_{(\lambda_1,\dots,\lambda_n)} \frac{\partial^{\lambda_1+\dots+\lambda_n}V}{\partial x_1^{\lambda_1}\dots\partial x_n^{\lambda_n}} \frac{\left(\frac{\hbar}{i}\right)^{\lambda_1+\dots+\lambda_n-1}}{\lambda_1!\dots\lambda_n!} \frac{\partial^{\lambda_1+\dots+\lambda_n}W}{\partial p_1^{\lambda_1}\dots\partial p_n^{\lambda_n}}.
\end{aligned}
$$

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Now let $W$ take $\hbar$ as a variable and consider a formal series expansion in $\hbar$. When $\hbar = 0$ we must get the classical result $e^{-\beta\left(V(\vec{x})+\sum_k \frac{p_k^2}{2m_k}\right)}$. For convenience, define $E = V(\vec{x}) + \sum_k \frac{p_k^2}{2m_k}$. Therefore we may write

$$
W(\vec{x}, \vec{p}) = e^{-\beta E} + \hbar W_1 + \hbar^2 W_2 + \hbar^3 W_3 + \dots. \tag{6.14}
$$

We may plug Equation 6.14 into Equation 6.13 and compare terms order-by-order in $\hbar$, since we expect $\frac{\partial W}{\partial t} = 0$ at thermal equilibrium since $W$ is in a stationary state. Doing so yields an infinite sequence of differential equations for all the $W_i$. These start with:

$$\sum_k -\frac{p_k}{m_k}\frac{\partial e^{-\beta E}}{\partial x_k} + \sum_k \frac{\partial V}{\partial x_k}\frac{\partial e^{-\beta E}}{\partial p_k} = 0$$

$$\sum_k -\frac{p_k}{m_k}\frac{\partial W_1}{\partial x_k} + \sum_k \frac{\partial V}{\partial x_k}\frac{\partial W_1}{\partial p_k} = 0 \qquad (6.15)$$

$$\sum_k -\frac{p_k}{m_k}\frac{\partial W_2}{\partial x_k} + \sum_k \frac{\partial V}{\partial x_k}\frac{\partial W_2}{\partial p_k} - \sum_k \frac{\hbar^2}{24}\frac{\partial^2 V}{\partial x_k^2}\frac{\partial^2 e^{-\beta E}}{\partial p_k^2} - \sum_{k\neq l}\frac{\hbar^2}{8}\frac{\partial^3 V}{\partial x_k^2 \partial x_l}\frac{\partial^3 e^{-\beta E}}{\partial p_k^2 \partial p_l} = 0$$

These are inhomogeneous differential equations in each of the $W_i$; the homogeneous part is always the same

$$\sum_k -\frac{p_k}{m_k}\frac{\partial W_i}{\partial x_k} + \sum_k \frac{\partial V}{\partial x_k}\frac{\partial W_i}{\partial p_k} = 0.$$

This is the differential equation for the stationary nature of the classical probability distribution, so its only solution is the classical probability distribution (or multiples thereof), else classical mechanics would be overdetermined. We've already accounted for multiples of the classical probability distribution in the constant term of the Wigner expansion, so we may as well ignore the homogeneous part [52].

The inhomogeneous part of the $W_1$ term's differential equation from Equation 6.15 is 0, so $W_1 = 0$. The inhomogeneous part of all the odd terms in the expansion contains $W_1$ and lower-order odd terms, so all odd terms in the Wigner expansion are 0 [52].

We can solve the $W_2$ term's differential equation from Equation 6.15 with a couple of clever substitutions and an integration. We find

$$W_2 = e^{-\beta E}\left[\sum_k\left(-\frac{\beta^2}{8m_k}\frac{\partial^2 V}{\partial x_k^2} + \frac{\beta^3}{24m_k}\left(\frac{\partial V}{\partial x_k}\right)^2\right) + \sum_{k,l}\frac{\beta^3 p_k p_l}{24m_k m_l}\frac{\partial^2 V}{\partial x_k \partial x_l}\right]. \qquad (6.16)$$

It is possible to solve for further terms in the Wigner expansion, but we will mostly not need that for our work ahead [13].

Thus far, we've considered the Wigner expansion as a formal series expansion, without claiming anything about convergence of the Wigner expansion anywhere. Understanding the convergence of the Wigner expansion in general or even in any particular specific case remains an open problem, and it is not clear that the Wigner expansion behaves particularly well. However, we expect the Wigner expansion to be well-behaved at higher temperatures or lower $\beta$, since that suppresses higher-order terms in the expansion. Further, empirical observations have established that the Wigner expansion, as opposed to any other semiclassical expansion of the Wigner function, has provided accurate approximations for use in simulation, so we chose to use the Wigner expansion to help us approximate the force-field functor [32].

## 6.2.2 Approximations to the Wigner Expansion

Assuming truncations of the Wigner expansion are a reasonable approximation to the full Wigner function, we may take the expansion to second-order, use Equation 6.16 to

compute the second-order correction, and then use Equation 6.12 to compute the new effective potential. But this effective potential would be extremely difficult to compute; in particular, while our original potential $V(\vec{x})$ is usually pairwise additive, $W(\vec{x}, \vec{p})$ is not and $\widetilde{V}(\vec{x})$ is not. Thus MD simulation algorithms would need to compute $N$-body terms, resulting in a runtime of $\mathcal{O}(N^N)$. Thus we will need to make further approximations using Taylor series expansions. Fortunately, it is easy to characterize, justify, and verify these assumptions for any force field.

Now suppose our potential is the sum of pairwise terms $V = \sum_{i,j} V_{ij}$ where $V_{ij}$ only depends on $x_i$ and $x_j$. Then we may rewrite Equation 6.16 as

$$
W(\vec{x}, \vec{p}) \approx e^{-\beta E(\vec{x}, \vec{p})} \left[ 1 + \hbar^2 \beta^2 \left( \sum_k \left( -\frac{1}{8m_k} \sum_j \frac{\partial^2 V_{jk}}{\partial x_k^2} + \frac{\beta}{24 m_k} \left( \sum_j \frac{\partial V_{jk}}{\partial x_k} \right)^2 \right) + \sum_{k,l} \frac{\beta p_k p_l}{24 m_k m_l} \frac{\partial^2 V_{kl}}{\partial x_k \partial x_l} \right) \right].
$$
(6.17)

The first and third terms in the second-order correction are immediately sums of the appropriate correction for each pairwise potential individually. The second term has the sums of each pairwise contribution, along with additional three-body terms of the form $\sum_{j,k,l} \frac{\beta}{12 m_k} \frac{\partial V_{jk}}{\partial x_k} \frac{\partial V_{kl}}{\partial x_k}$. We will see that for the potentials we will test our methods on these terms are often smaller than the others in the Wigner expansion by a significant margin since first derivatives are smaller than second derivatives, so we will ignore the non-additive terms and assume that the Wigner expansion is also pairwise additive for our purposes. If this assumption does not hold, it is simply an additional force to be added. For convenience, we write

$$
W(\vec{x}, \vec{p}) = e^{-\beta E(\vec{x}, \vec{p})} \left[ 1 + \hbar^2 \beta^2 \sum_{j,k} W_{jk} + \dots \right]
$$
(6.18)

where $W_{jk}$ corresponds to the second-order Wigner correction from $V_{jk}$.

Now let us compute the effective potential. We have

$$
\widetilde{V}(\vec{r}) = -\frac{1}{\beta} \log \left[ \int d\vec{p} \, e^{-\beta E(\vec{x}, \vec{p})} \left( 1 + \hbar^2 \beta^2 \sum_{j,k} W_{jk} + \dots \right) \right]
$$

$$
= -\frac{1}{\beta} \left( \log \left[ \int d\vec{p} \, e^{-\beta E(\vec{r}, \vec{p})} \right] + \log \left[ 1 + \frac{\hbar^2 \beta^2 \sum_{j,k} \int d\vec{p} \, e^{-\beta E(\vec{r}, \vec{p})} W_{jk}}{\int d\vec{p} \, e^{-\beta E(\vec{x}, \vec{p})}} \right] \right)
$$

$$
\approx -\frac{1}{\beta} \left( \log \left[ \int d\vec{p} \, e^{-\beta E(\vec{x}, \vec{p})} \right] + \sum_{j,k} \log \left[ 1 + \hbar^2 \beta^2 \frac{\int d\vec{p} \, e^{-\beta E(\vec{x}, \vec{p})} W_{jk}}{\int d\vec{p} \, e^{-\beta E(\vec{r}, \vec{p})}} \right] \right)
$$

where we use the first-order Taylor series approximation $\log\left( 1 + \sum_i x_i \right) \approx \sum_i \log(1 + x_i)$. This approximation is valid anywhere the correction from the Wigner expansion is relatively small, i.e. at high enough temperatures such that the Wigner expansion itself is valid. The first term is the classical potential up to normalization and the second term is

the pairwise sum of the individual contributions to the effective potential. Thus under the two approximations we have made, we see that the $N$-body problem reduces successfully to a pairwise sum of 2-body problems.

The calculations done here can be trivially generalized to potentials that are the sums of multi-body interactions (like bond angles or torsions). Thus we must check that the approximations made in this section are still valid for the more complicated potential; if they are, then our method can be applied by simply computing the correction to each individual term using Equation 6.16. Even if our first assumption of the product-of-first-derivative terms being negligible fails, we can simply add those terms in; derivatives are generally short-range, so these terms would continue to be easy to compute. We have not yet found a way to generalize this method to potentials that are not pairwise additive.

### 6.2.3 Empirical Verification on Neon

We proceed to test the Wigner expansion approximation to the force-field functor on a Neon force field. Neon is spherically symmetric, allowing for convenient computations. Specifically, we found that for a spherically symmetric system, the non-normalized correction term is

$$W_2(r, p, \theta) = \frac{e^{-\beta E(r,p,\theta)}}{24m^2 r} \left( -6m \frac{\partial V}{\partial r} + \beta p^2 \sin^2 \theta \frac{\partial V}{\partial r} + mr\beta \left( \frac{\partial V}{\partial r} \right)^2 - 3mr \frac{\partial^2 V}{\partial r^2} + \beta p^2 r \cos^2 \theta \frac{\partial^2 V}{\partial r^2} \right).$$

(6.19)

Here, $r$ is radial distance, $p$ is the magnitude of the momentum, and $\theta$ is the angle between position and momentum.

We test our method on a Neon force field. Neon's interatomic potential is spherically symmetric and a highly accurate force field for Neon was derived by [28] by fitting functional forms to high-quality CCSD(T) electronic structure data.

We began by numerically computing the effective potential at various temperatures as shown in Figure 6.2.1a. We note two characteristics of the effective potential which can be explained by quantum phenomena: First, the repulsive wall as $r \to 0$ is less steep in the effective potential than in the classical potential. This is due to the fact that the quantum wavepacket can enter the classically forbidden region $E < V(r)$. Second, the well is shallower, due to the zero-point energy.

Moreover, the equilibrium distance is slightly longer in the effective potential than in the classical potential. As we show below, this has important thermodynamic consequences.

We note a bump in the effective force field at $T = 25 \, \text{K}$. This is likely an artifact due to use of only the second-order expansion and neglecting the $N$-body terms, and not a reflection of the full quantum effective force field. This bump is significantly larger in the effective force than in the potential. The fourth-order correction does reduce the bump in the repulsive wall while not significantly affecting the location or depth of equilibrium length and thus shows an improvement over the second-order term; a graph is shown in Figure 6.2.1b.
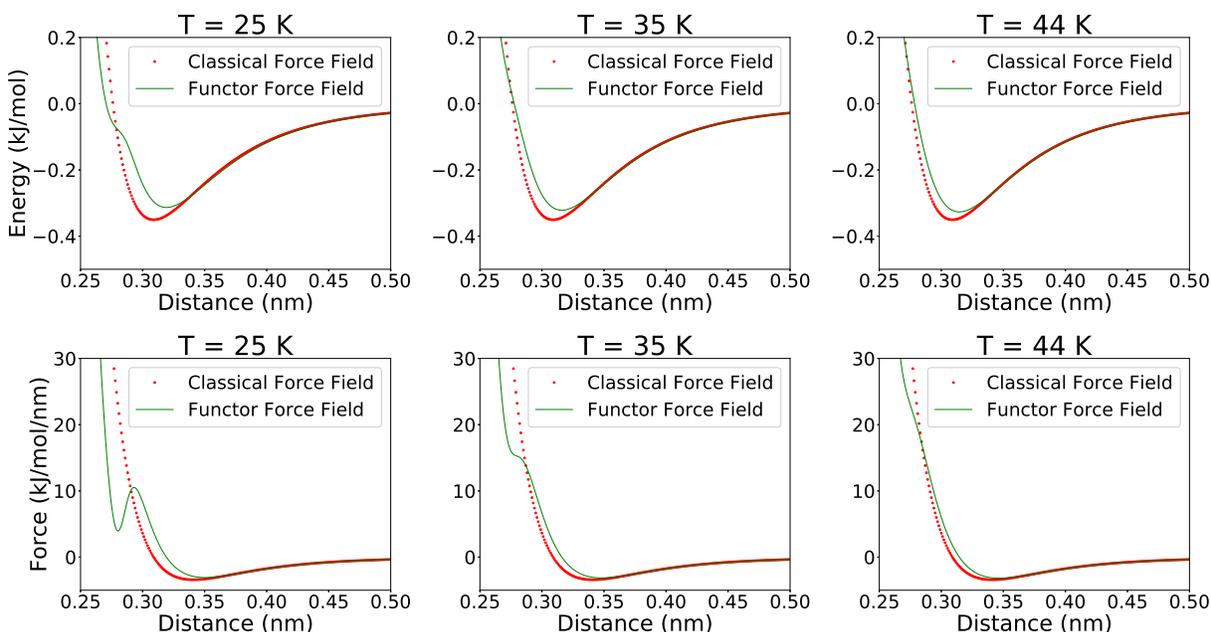
(a) Second-Order Corrections.
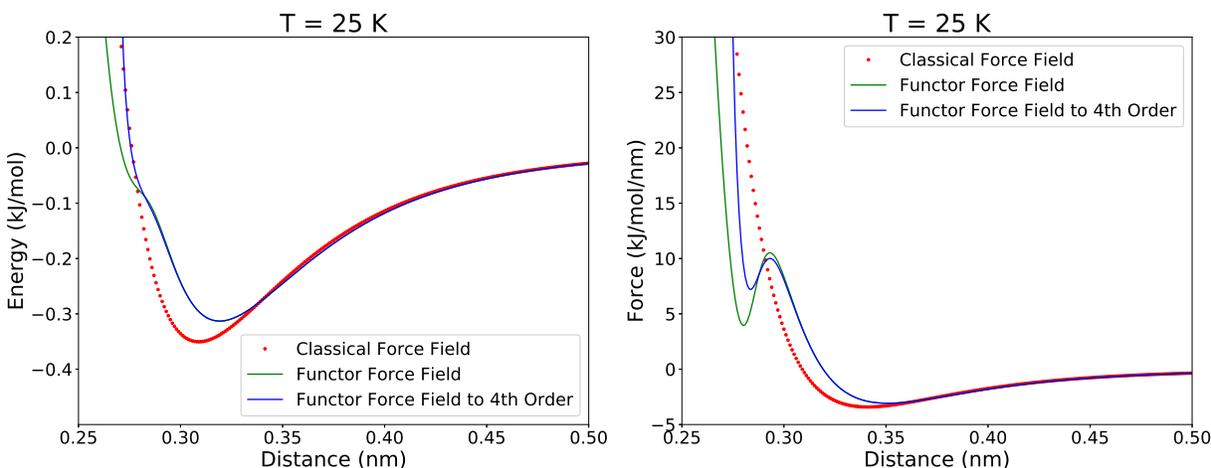


(b) Fourth-Order Correction.

Figure 6.2.1: Classical and effective force field for Neon at $T = 25\,\mathrm{K}, 35\,\mathrm{K}$, and $T = 44\,\mathrm{K}$. The red dotted curve is the classical potential from Hellmann et al. [28], the green curve is the effective potential numerically computed using the Wigner expansion to second order, and the blue curve adds corrections to fourth order on the bottom for $T = 25\,\mathrm{K}$. The change in shape and shallower well can be explained by the wavepacket entering the classically forbidden region and the zero-point energy, respectively. The fourth-order correction reduces the bump in the repulsive wall, thus showing an improvement from the second-order.

### 6.2.3.1   Radial Distribution Function

The first parameter we chose to measure in simulation was the radial distribution function (RDF). Given a system of $N$ liquid or gas spherically symmetric particles in a volume $V$, the RDF $g(r)$ is defined as

$$n(r) = 4\pi r^2 \rho g(r)\, dr \tag{6.20}$$

where $n(r)$ is the number of particles in a spherical shell of radius $r$ and thickness $dr$ and $\rho = \frac{N}{V}$ is the average number density. $g(r)$ thus represents the correlations between particle positions as a result of the interparticle potential; should there be no potential (as in the approximation of an ideal gas), then the density function would simply be $4\pi r^2 \rho\, dr$. The RDF is strongly potential-dependent and thus a good initial candidate for an experimental observable to use in verifying our force-field functor.

Computing the RDF from a simulation is also straightforward, since it is a purely position-dependent observable. We carry out the number-counting suggested in Equation 6.20 to compute $n(r)$ for some fixed, very small $dr$, and then compute $g(r)$ using Equation 6.20. The result is then averaged across all frames.

We ran NPT simulations with a Monte Carlo barostat to fix the pressure; each time step was 2 fs. The initial condition was a cubic lattice with atoms spaced 0.3 nm apart to emulate the density of the liquid phase. Since all experimental measurements were done on liquid Neon, we chose to begin all simulations with an approximately liquid-phase initial condition. We ran the simulation for 100 000 steps (using the initial 10 000 for equilibration) and verified that the simulations converged by checking the computed density.

For verification of our results, we compared both with experimental results from Bellissent-Funel et al. [6] and de Graaf and Mozer [14], purely classical MD, and an alternative simulation method, Ring Polymer Molecular Dynamics (RPMD), that accounts for nuclear quantum effects by modifying the underlying MD algorithm itself. We omit the details of how RPMD works here; they can be found in Allen and Tildesley [1]. RPMD simulations replace every atom with $P$ beads, or copies, and run molecular dynamics; we chose $P = 32$. The computational time complexity of RPMD is $\mathcal{O}(N^2 P)$, so we see a 32 times slowdown using RPMD, as opposed to the force-field functor theory method. We will see that using force-field functor theory causes no loss in accuracy as compared to RPMD. Figure 6.2.2 shows radial distribution functions (RDF) for liquid Neon at $T = 26.1$ K, $T = 35.05$ K, and $T = 42.2$ K. The classical MD simulation with a Born-Oppenheimer force field will be called C-FF for short and the classical MD simulation with our effective force field will be called E-FF.

At $T = 35.05$ K and $T = 42.2$ K, E-FF is roughly as accurate as RPMD, with both being a significant improvement relative to C-FF. In particular, the location and height of the first peak obtained by E-FF are more precise than those derived by C-FF. The shift and lower height of the peak are caused by quantum corrections which shift the equilibrium position and reduce the depth of the well in the effective potential described above.

At $T = 26.1$ K, E-FF, while still a significant improvement over C-FF, is somewhat worse than RPMD. Specifically, the location and height of the peak are relatively accurate, but the bump in the potential noted above results in a similar bump in the RDF. Inaccuracies in both RPMD and E-FF at this temperature indicate more beads and higher-order
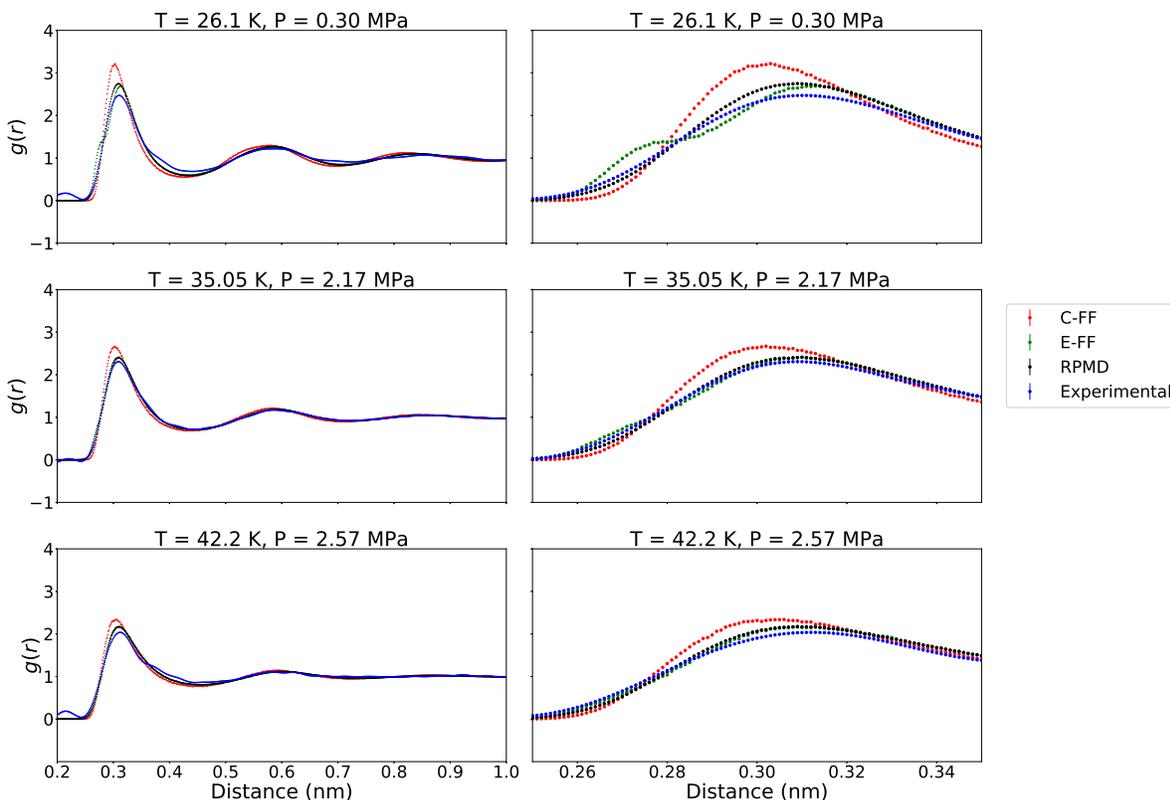
Figure 6.2.2: Radial distribution function as computed by C-FF, E-FF, and RPMD. In each plot, the red curve is computed by C-FF, the green curve is computed by E-FF, the black curve is computed by RPMD, and the blue curve is experimentally determined from [6] and [14]. The right side focuses on just the first peak. In all cases, the height and location of the first peak are more accurately determined by E-FF and RPMD than by C-FF. E-FF demonstrates similar accuracy as the significantly slower RPMD.

terms are necessary for an accurate computation of the RDF. We verified that a 4th-order computation of the RDF reduced the size of the bump at $T = 26.1$ K.

### 6.2.3.2 Equation of State

The next parameter we chose to compute was the equation of state for Neon, or the relationship between density and pressure for a variety of temperatures. We used similar NPT simulations as we did in measuring the radial distribution function. The initial condition in all cases was a cubic lattice with atoms spaced 3 nm apart to emulate the density of the gas phase. Experimental measurements were done in both liquid and gas phase; we ran the simulation for 1 000 000 steps (using the initial 100 000 for equilibration) and verified convergence by checking the density values.

Figure 6.2.3 shows the computed equation of state for Neon at $T = 36$ K and $T = 40$ K. Experimental data [21] is compared to densities for both the classical and effective potential [17]. These values span liquid and gas phase.

Both E-FF and C-FF were accurate on gas-phase densities. The maximum error relative
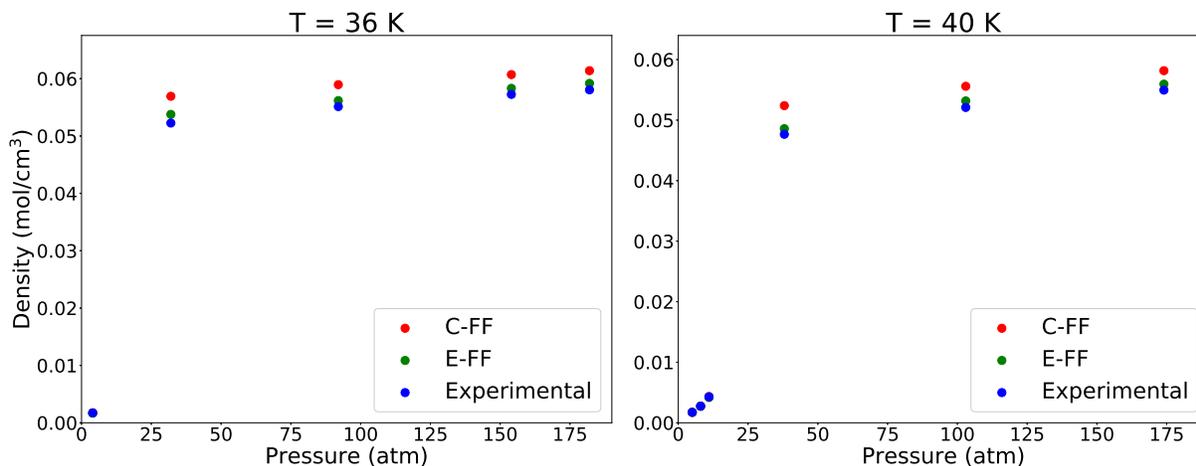
Figure 6.2.3: Equation of state as computed by C-FF and E-FF at $T = 36\,\text{K}, 40\,\text{K}$. The red points were computed by C-FF, the green points by E-FF, and the blue points experimentally determined by [21]. The gas-phase density points from all three methods are too close to be distinguished. E-FF is significantly more accurate in liquid-phase, reducing the error relative to experimental data from 10% for C-FF to 2% for E-FF.

to the experimental data for C-FF is around 1% and for E-FF around 2%. This is somewhat expected since gas-phase densities do not depend much on the force field. In liquid phase, C-FF significantly overestimates the experimental density by as much as 10% while E-FF only overestimates the experimental density by around 2%. This is due to the shift in equilibrium position described above. Adding additional terms to the Wigner expansion should further reduce the error in calculated density.

### 6.2.3.3 Vapor-Liquid Coexistence Curve

The last parameter we calculated was the vapor-liquid coexistence curve, or the temperature/pressure combination at which gas and liquid phase coexist. At atmospheric pressure, this point is known as the boiling point. This is an observable that cannot be easily mathematically related to a function of position and momentum, unlike the density or RDF. Verifying the accuracy of the force-field functor on the vapor-liquid coexistence curve suggests broader applicability of the functor beyond the observables directly guaranteed to include nuclear quantum effects by Theorem 6.1.1. In this section, we assume some background in the statistical mechanics of phase transitions, since introducing that material would take us too far afield.

Computing the vapor-liquid coexistence curve cannot be done reliably by MD simulation. We therefore used Gibbs Monte Carlo (MC) simulations instead [44, 43, 38]. Normal MC perturbation steps are performed as in a typical NVT simulation, but we also randomly transfer volume and particles between the two boxes to equilibrate pressure and chemical potential. By the definition of a phase transition (known as the Gibbs condition), both boxes will end up on the vapor-liquid coexistence curve. We then measured density and pressure (from the virial theorem [1]) to compute the phase transition points.

This simulation was implemented on MCCCS-Towhee [37]. Far from the critical point,

(a) Density vs. Temperature

(b) Temperature vs. Pressure

(c) Absolute Density Error vs. Density

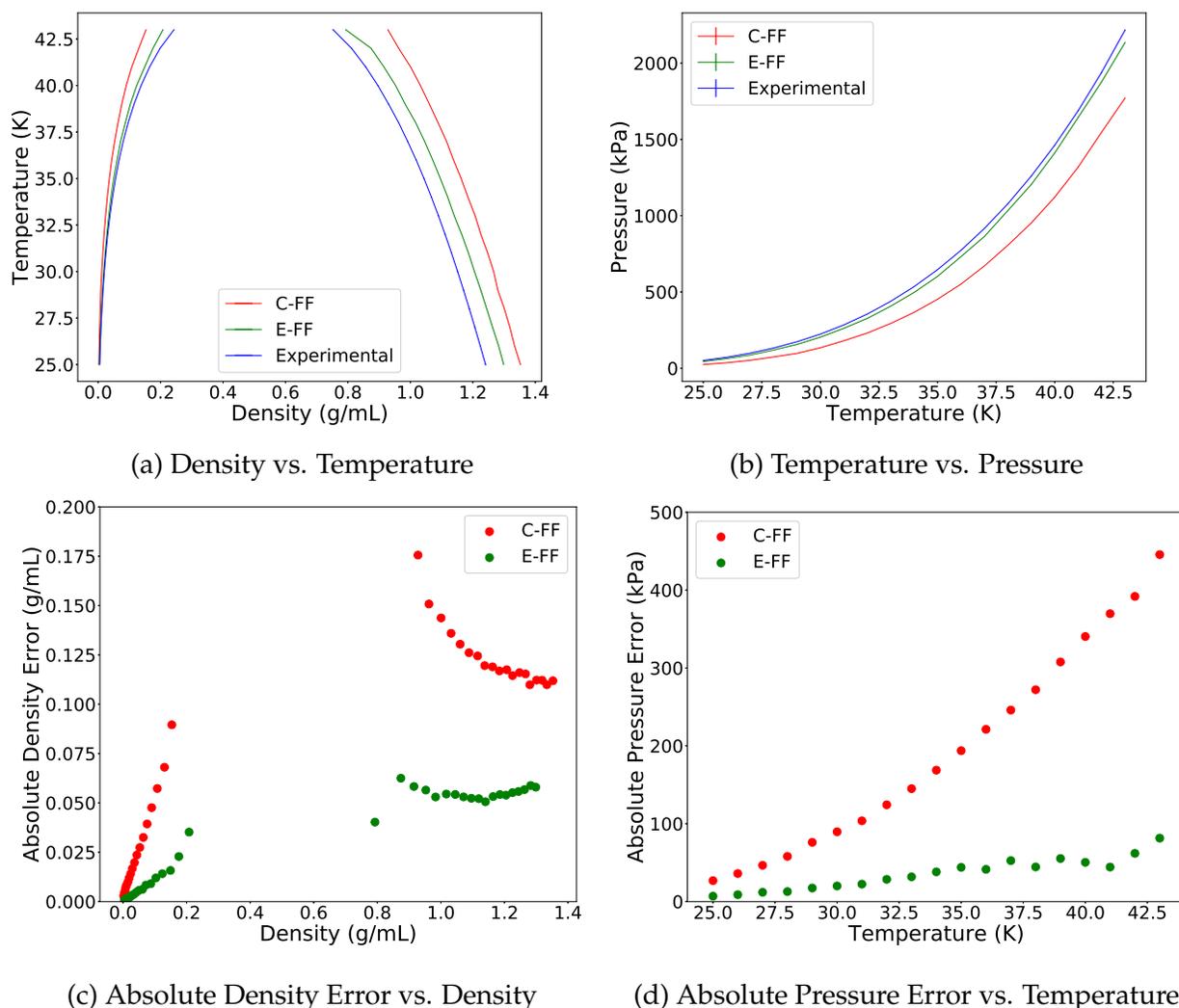(d) Absolute Pressure Error vs. Temperature

Figure 6.2.4: Vapor-liquid coexistence curve for Neon, as computed by Gibbs MC simulation with both C-FF and E-FF. The red curves are computed by C-FF, the green curves by E-FF, and the blue curves experimentally determined by [49]. E-FF is significantly more accurate in all cases. It reduces the average error relative to the experimental data: from 40% to 11% for the gas phase density; from 12% to 5% for the liquid phase density; from 33% to 7% for the pressure.

we followed the guidelines of Morales et al. [38] in setting up the simulation. We began with one empty box and one box full of 1000 Neon atoms; both boxes were cubes with side length 4 nm. At the end of the simulation, around 10% of the atoms were in vapor phase, as suggested by Morales et al. [38] for the most rapid convergence. We set the probability of a perturbation to the default value and ran the simulation for 10 000 cycles, sufficient to guarantee convergence. We verified convergence by checking that the density values converged appropriately.

It is known that Gibbs MC simulations fail close to the critical point [15] since the

two phases are very similar, causing the boxes to interconvert between phase frequently. At $T = 43$ K, we adjusted our analysis for the effective potential simulation to use a density probability distribution to identify liquid and vapor phase densities and ignore all intermediate phases [15]. We did not analyze $T = 44$ K since it was too close to the critical point for the effective potential simulation. The classical simulation did not exhibit any behavior characteristic of the critical point, so no adjustments were needed.

E-FF is significantly more accurate in determining both the density and the pressure of the vapor-liquid coexistence point, as shown in Figure 6.2.4. Specifically, C-FF consistently overestimates the liquid density by an average percent error of 12% relative to experimental data, while E-FF consistently overestimates it by only 5%. This effect can be partly explained by the density deviations observed when computing the equation of state. C-FF underestimates gas density by an average percent error of 40% relative to the experimental data, while E-FF underestimates it by 11%. Finally, C-FF underestimates the pressure by an average percent error of 33% relative to the experimental data while E-FF underestimates it by 7%. Larger errors at lower temperature suggest that the effective force field would be improved by using higher-order corrections.

We also note that E-FF began exhibiting critical behavior, i.e., transitions between vapor and liquid phase, at around $T = 43$ K, very close to the experimental critical point of $T_c = 44.40$ K. C-FF did not exhibit similar behavior even at $T = 44$ K. We chose not to measure the critical point directly due to inaccuracies with Gibbs MC simulations close to the critical point [15], but this result suggests that C-FF also overestimates the critical point.

Our empirical work with Neon, our analytic work with the Wigner expansion, and the fundamental results of force-field functor theory makes clear that we have a new way of accurately and efficiently accounting for nuclear quantum effects. Further, the work we have done above shows that generalizing these methods to more complicated force fields should be possible, provided the assumptions that we have outlined above hold for those more complicated force fields. Application of this method would significantly improve the accuracy of most MD simulations by avoiding errors associated with nuclear quantum effects as outlined above.

# Chapter 7

# Conclusions and Future Work

Fully understanding the errors associated with molecular dynamics simulations is certainly a work-in-progress. While bounds have been derived to characterize almost every error associated with an MD simulation, these bounds are in general either rather weak or require additional conditions beyond those that can be assumed for typical simulations. Here, we outline some of the remaining questions raised by the results presented in this thesis:

1. Can we bound time discretization and sampling error while dropping the integrability condition used in Theorems 3.2.2 and 4.2.2? Alternatively, can we provide some rigorous justification for why MD force fields are integrable or near-integrable? The justification provided thus far is just a conjecture.

2. Is it possible to exactly fit a force field to the lowest-energy Born-Oppenheimer electronic state? This question is the driving source for almost all current force field development research.

3. Following the above, what is the resulting error in the measurement of any observable given a certain error in the force field? Understanding this would help us set goals on how accurate our force fields needed to be to carry out any given measurement.

4. Does force-field functor theory work in full generality for any force field? Further empirical testing and analytic results on the Wigner expansion are needed to understand the merits of this approach. If not, is there a better way of computing the force-field functor without the Wigner expansion? Ideally, we would like a method that worked at all temperatures and for which mathematical convergence was known.

Answering these questions will help scientists better understand the error in the MD simulations they run. Such bounds will make biochemists more confident in drug predictions made via simulation or material scientists more confident in designing materials with properties predicted via simulation. If we understand the size of errors in observables predicted by MD simulation, we can be more confident in results predicted by simulation and thus more likely to use simulations to answer major scientific questions.

# Bibliography

[1] M.P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Clarendon Press, Oxford, UK, 1987.

[2] V. I. Arnold. *Mathematical Methods of Classical Mechanics*. Springer, New York, second edition, 1989.

[3] Ryan Babbush, John Parkhill, and Alán Aspuru-Guzik. Force-Field Functor Theory: Classical Force Fields Which Reproduce Equilibrium Quantum Distributions. *Front. Chem.*, 1:26, 2013.

[4] C. Bayer, H. Hoel, P. Plechac, A. Szepessy, and R. Tempone. How Accurate is Molecular Dynamics? *ArXiV e-prints*, 2011.

[5] Christian Bayer, Hakon Hoel, Ashraful Kadir, Petr Plechac, Mattias Sandberg, and Anders Szepessy. Computational Error Estimates for Born-Oppenheimer Molecular Dynamics with Nearly Crossing Potential Surfaces. *Applied Mathematics Research eXpress*, 2015(2):329 – 417, 2015.

[6] M. C. Bellissent-Funel, U. Buontempo, A. Filabozzi, C. Petrillo, and F. P. Ricci. Neutron Diffraction of Liquid Neon and Xenon Along the Coexistence Line. *Phys. Rev. B*, 45(9):4605–4613, March 1992.

[7] Stephen D. Bend and Benedict J. Leimkuhler. Molecular Dynamics and the Accuracy of Numerically Computed Averages. *Acta Numer.*, pages 1 – 65, 2007.

[8] E. Cances, F. Castella, P. Chartier, E. Faou, C. Le Bris, F. Legoll, and G. Turinci. High-Order Averaging Schemes with Error Bounds for Thermodynamical Properties Calculations by Molecular Dynamics Simulations. *J. Chem. Phys.*, 121(21), December 2004.

[9] E. Cances, F. Castella, P. Chartier, E. Faou, C. Le Bris, F. Legoll, and G. Turinci. Long-Time Averaging for Integrable Hamiltonian Dynamics. *Numer. Math.*, 100:211 – 232, 2005.

[10] William B. Case. Wigner Functions and Weyl Transforms for Pedestrians. *Am. J. Phys.*, 76(10):937 – 946, October 2008.

[11] Michele Ceriotti, David E. Manolopoulos, and Michele Parrinello. Accelerating the Convergence of Path Integral Dynamics with a Generalized Langevin Equation. *J. Chem. Phys.*, 134(084104), 2011.

[12] David Chandler and Peter G. Wolynes. Exploiting the Isomorphism between Quantum Theory and Classical Statistical Mechanics of Polyatomic Fluids. *J. Chem. Phys.*, 74:4078 – 4095, 1981.

[13] W. T. Coffey, Yu. P. Kalmykov, S. V. Titov, and B. P. Mulligan. Wigner Function Approach to the Quantum Brownian Motion of a Particle in a Potential. *Phys. Chem. Chem. Phys.*, 9:3361–3382, 2007.

[14] L. A. de Graaf and B. Mozer. Structure Study of Liquid Neon by Neutron Diffraction. *J. Chem. Phys.*, 55(10):4967–4973, November 1971.

[15] Mohammadhasan Dinpajooh, Peng Bai, Douglas A. Allan, and J. Ilja Siepmann. Accurate and Precise Determination of Critical Properties from Gibbs Ensemble Monte Carlo Simulations. *J. Chem. Phys.*, 143(114113), 2015.

[16] Ron O. Dror, Robert M. Dirks, J. P. Grossman, Huafeng Xu, and David E. Shaw. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu. Rev. Biophys.*, 41:429 – 452, 2012.

[17] P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L.-P. Wang, D. Shukla, M. Houston, T. Stitch, C. Klein, M. R. Shirts, and V. S. Pande. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J. Chem. Theory Comput.*, 9(1): 461–469, 2013.

[18] Gérard G. Emch and Chuang Liu. *The Logic of Thermostatistical Physics*. Springer-Verlag, New York, 2002.

[19] George S. Fanourgakis, Thomas E. Markland, and David E. Manolopoulos. A Fast Path Integral Method for Polarizable Force Fields. *J. Chem. Phys.*, 131(094102), 2009.

[20] L. D. Fosdick and H. F. Jordan. Path Integral Calculation of the Two Particle Slater Sum for Helium. *Phys. Rev. B*, 143:58 – 66, 1966.

[21] R. M. Gibbons. The Equation of State of Neon Between 27 and 70 K. *Cryogenics*, pages 251–260, August 1969.

[22] David J. Griffiths. *Introduction to Quantum Mechanics*. Pearson Prentice Hall, Upper Saddle River, NJ, 2004.

[23] Stephen F. Gustafson and Israel Michael Sigal. *Mathematical Concepts of Quantum Mechanics*. Springer-Verlag, Berlin Heidelberg, 2003.

[24] Scott Habershon, David E. Manolopoulos, Thomas E. Markland, and Thomas F. Miller. Ring-Polymer Molecular Dynamics: Quantum Effects in Chemical Dynamics from Classical Trajectories in an Extended Phase Space. *Annu. Rev. Phys. Chem.*, 64: 387 – 413, 2013.

[25] E. Hairer, C. Lubich, and G. Wanner. Geometric Numerical Integration Illustrated by the St ormer-Verlet Method. *Acta Numer.*, pages 399 – 450, 2003.

[26] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer, New York, second edition, 2006.

[27] Eric J. Heller. *The Semiclassical Way to Dynamics and Spectroscopy*. Unpublished.

[28] Robert Hellmann, Eckard Bich, and Eckhard Vogel. *Ab Initio* Potential Energy Curve for the Neon Atom Pair and Thermophysical Properties of the Dilute Neon Gas. I. Neon-Neon Interatomic Potential and Rovibrational Spectra. *Mol. Phys.*, 106(1):133–140, January 2008.

[29] P. D. Hislop and I. M. Sigal. *Introduction to Spectral Theory: With Applications to Schrödinger Operators*. Springer-Verlag, 1996.

[30] Seogjoo Jang, Soonmin Jang, and Gregory A. Voth. Applications of Higher Order Composite Factorization Schemes in Imaginary Time Path Integral Simulations. *J. Chem. Phys.*, 115:7832 – 7842, 2001.

[31] C. Le Bris. Computational Chemistry from the Perspective of Numerical Analysis. *Acta Numer.*, pages 363 – 444, May 2005.

[32] Hai-Woong Lee. Theory and Application of the Quantum Phase-Space Distribution Functions. *Phys. Rep.*, 259:147 – 211, 1995.

[33] Ben Leimkuhler and Charles Matthews. *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*. Springer, London, 2015.

[34] Jian Liu, William H. Miller, Francesco Paesani, Wei Zhang, and David A. Case. Quantum Dynamical Effects in Liquid Water: A Semiclassical Study on the Diffusion and the Infrared Absorption Spectrum. *J. Chem. Phys.*, 131(164509), 2009.

[35] Pedro E. M. Lopes, Olgun Guvench, and Jr. Alexander D. MacKerell. Current Status of Protein Force Fields for Molecular Dynamics. *Methods Mol. Biol.*, 1215:47 – 71, 2015.

[36] L. Markus and K. R. Meyer. Generic Hamiltonian Dynamical Systems are Neither Integrable nor Ergodic. *Mem. Amer. Math. Soc.*, 144:1 – 57, 1974.

[37] Marcus G. Martin. MCCCS Towhee: A Tool for Monte Carlo Molecular Simulation. *Mol. Simul.*, pages 1212 – 1222, 2013.

[38] Angel D. Cortes Morales, Ioannis G. Economou, Cornelis J. Peters, and J. Ilja Siepmann. Influence of Simulation Protocols on the Efficiency of Gibbs Ensemble Monte Carlo Simulations. *Mol. Simul.*, pages 1135 – 1142, 2013.

[39] Cameron Mura and Charles E. McAnany. An Introduction to Biomolecular Simulations and Docking. *Mol. Simul.*, 40(10):732 – 764, 2014.

[40] Akira Nakayama and Nancy Makri. Forward-Backward Semiclassical Dynamics for Quantum Fluids using Pair Propagators: Application to Liquid *para*-Hydrogen. *J. Chem. Phys.*, 119(16):8592 – 8605, October 2003.

[41] Dion Robert James O'Neale. *Preservation of Phase Space Structure in Symplectic Integration*. PhD thesis, Massey University, 2009.

[42] Francesco Paesani, Wei Zhang, David A. Case, Thomas E. Cheatham III, and Gregory A. Voth. An Accurate and Simple Quantum Model for Liquid Water. *J. Chem. Phys.*, 125(184507), 2006.

[43] A. Z. Panagiotopoulos, N. Quirke, M. Stapleton, and D. J. Tildesley. Phase Equilibria by Simulation in the Gibbs Ensemble. *Mol. Phys.*, 63(4):527–545, 1988.

[44] Athanassios Z. Panagiotopoulos. Direct Determination of Phase Coexistence Properties of Fluids by Monte Carlo Simulation in a New Ensemble. *Mol. Phys.*, 61(4): 813–826, 1987.

[45] Marc Pavese, Soonmin Jang, and Gregory A Voth. Centroid Molecular Dynamics: A Quantum Dynamics Method Suitable for the Parallel Computer. *Parallel Computing*, 26:1025 – 1041, 2000.

[46] Álvaro Pelayo. Hamiltonian and Symplectic Symmetries: An Introduction. *ArXiV e-Prints*, 2016.

[47] Alejandro Pérez and Mark E. Tuckerman. Improving the Convergence of Closed and Open Path Integral Molecular Dynamics via Higher Order Trotter Factorization Schemes. *J. Chem. Phys.*, 135(064104), 2011.

[48] Jens Aage Poulsen, Gunnar Nyman, and Peter J. Rossky. Practical Evaluation of Condensed-Phase Quantum Correlation Functions: A Feynman-Kleinert Variational Linearized Path Integral Method. *J. Chem. Phys.*, 119(23), 2003.

[49] V. A. Rabinovich, A. A. Vasserman, V. I. Nedostup, and L. S. Veksler. *Thermophysical Properties of Neon, Argon, Krypton, and Xenon*, volume 10 of *National Standard Reference Data Service of the USSR: A Series of Property Tables*. Hemisphere Publishing Corporation, 1988.

[50] Thomas Spura, Christopher John, Scott Habershon, and Thomas D. Kuhne. Nuclear Quantum Effects in Liquid Water from Path-Integral Simulations using an *Ab Initio* Force Matching Approach. *Mol. Phys.*, 113(8):1 – 15, 2015.

[51] Vikram Sundar, David Gelbwaser-Klimovsky, and Alán Aspuru-Guzik. Reproducing Quantum Probability Distributions at the Speed of Classical Dynamics: A New Approach for Developing Force-Field Functors. *J. Phys. Chem. Lett.*, 2018. doi: 10.1021/acs.jpclett.7b03254. In press.

[52] E. Wigner. On the Quantum Correction for Thermodynamic Equilibrium. *Phys. Rev.*, 40:749–759, 1932.