SYMPLECTIC GEOMETRY AND HAMILTONIAN MONTE CARLO METHOD

by

Feyza Öztürk

B.S., Mathematics, Boğaziçi University, 2019

Submitted to Kandilli Observatory and Earthquake
Research Institute in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Geophysics Department

Boğaziçi University

2022

SYMPLECTIC GEOMETRY AND HAMILTONIAN MONTE CARLO METHOD

DATE OF APPROVAL: 19.07.2022

To my beloved son, Odi ...

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Çağrı Diner who guided me in doing this thesis. He provided me with invaluable advice and helped me in my difficult periods. His motivation and help contributed tremendously to the successful completion of the thesis. Besides, I owe a big thank to Tevfik Mustafa Aktar as the second reader of this thesis. He provided me with continuous encouragement and helped me by giving support for computational part which I needed for my thesis.

I would like to express my special thanks to Ali Özgün Konca and Hayrullah Karabulut for sharing their knowledge and experiences.

Also, I would like to thank my family and friends for their emotional support. Without that support I could not have succeeded in completing this thesis.

At last but not in least, I would like to thank all who helped and motivated us to work on this project.

ABSTRACT

SYMPLECTIC GEOMETRY AND HAMILTONIAN MONTE CARLO METHOD

Hamiltonian Monte Carlo (HMC) method is an application of a non-Euclidean geometry to an inverse problem. HMC is a probabilistic sampling method with the basis of Hamiltonian dynamics. One of the main advantages of HMC algorithm is to draw independent samples from the model space with a higher acceptance rate than other Markov Chain Monte Carlo (MCMC) methods. In order to understand how higher acceptance rate is achieved, I have studied HMC in the light of symplectic geometry. Hamiltonian dynamics is defined on the phase space (cotangent bundle), which has a natural symplectic structure, i.e. a differential two-form which is non-degenerate and closed.

Hamiltonian function is defined on the phase space, which corresponds to the sum of misfit and the square of the generalized momentum. By using the non-degeneracy property of symplectic form, a vector field can be found in which Hamiltonian function is invariant along the integral curves of the vector field. The invariance of the Hamiltonian function results in high acceptance rate, where we apply accept-reject test to satisfy detailed-balance property.

In this thesis, we define some basic concepts and theorems in symplectic geometry, then describe the relation between symplectic geometry and HMC, namely Hamiltonian dynamics. Lastly, we show an implementation for HMC algorithm to a 2D-tomography problem and analyze the tune parameters for application of HMC.

ÖZET

SİMPLEKTİK GEOMETRİ VE HAMILTONIAN MONTE CARLO METODU

Hamilton Monte Carlo (HMC) yöntemi, Öklidyen olmayan bir geometrinin ters-çözüm problemlerine uygulanmasıdır; Hamilton dinamiğine dayanan olasılıksal bir örnekleme yöntemidir. HMC algoritmasının temel avantajlarından biri, diğer Monte Carlo Markov Zinciri yöntemlerine göre daha yüksek bir kabul oranı sahip ve bağımsız örnekler çizmesidir. Daha yüksek bir kabul oranının nasıl elde edildiğini anlamak için, simplektik geometri ışığında HMC metodunu inceledim. Hamilton dinamiği, dejenere olmayan ve kapalı bir diferansiyel 2-forma (simplektik form) sahip olan faz uzayında (kotanjant demeti) tanımlanır.

Hamilton fonksiyonu, gözlemlenen data ile tahmini data arasındaki fark toplamına ve genelleştirilmiş momentumun karesine karşılık gelir ve faz uzayında tanımlıdır. Simplektik formun dejenere olmama özelliğini kullanarak, Hamilton fonksiyonunun vektör alanının integral eğrileri boyunca değişmez olduğu bir vektör alanı bulunabilir. Hamilton fonksiyonunun değişmezliği, ayrıntılı denge özelliğini sağlamak için kabul-red testi uyguladığımız yüksek kabul oranı ile sonuçlanır.

Tezimde, önce simplektik geometrideki bazı temel kavramları ve teoremleri tanımlayacağız, ardından simplektik geometri ile HMC arasındaki ilişkiyi yani Hamilton dinamiğini anlatacağım. Son olarak, bir 2D tomografi problemi için HMC algoritmasının uygulanışını göstereceğim ve ayar parametrelerini analiz edeceğim.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

H	Hamilton function
K	Kinetic energy
L	Lagrange function
$\boldsymbol{p}$	Generalized momentum
T	Transition matrix
U	Potential energy
$\boldsymbol{q}$	Generalized position
Σ	Covariance matrix
$\mathcal{C}$	Configuration space *or* state space
$\mathcal{L}_{\boldsymbol{X}_H}(\cdot)$	Lie derivative along the vector field $\boldsymbol{X}_H$
μ	Mean
$\boldsymbol{\omega}$	Symplectic two-form
σ^2	Variance
$\otimes$	Outer product
$\lvert \cdot \rvert$	Euclidean norm
$\wedge$	Wedge product

LIST OF ACRONYMS/ABBREVIATIONS

2D	Two dimensional
$A[(\boldsymbol{q}_i, \boldsymbol{p}_i)]$	Acceptance probability of the proposed state $(\boldsymbol{q}_i, \boldsymbol{p}_i)$
gcd	Greatest common divisor
HMC	Hamiltonian Monte Carlo
MCMC	Markov Chain Monte Carlo
$T_{\boldsymbol{q}}\mathbf{R}^n$	Tangent Space to $\mathbf{R}^n$ at $\boldsymbol{q}$
$T^*_{\boldsymbol{q}}\mathbf{R}^n$	Cotangent Space to $\mathbf{R}^n$ at $\boldsymbol{q}$
$T\mathbf{R}^n$	Tangent Bundle to $\mathbf{R}^n$
$T^*\mathbf{R}^n$	Cotangent Bundle to $\mathbf{R}^n$
$\boldsymbol{X}_H$	Hamiltonian vector field

1. INTRODUCTION

Geophysics is mainly based on inverting a non-observable physical property from an observed data, e.g. velocity structure of the subsurface from the travel time data, earthquake location from traveltime data, fault position and its properties from earthquakes. Backus and Gilbert [1,2] had contributed to the establishment of inverse theory by combining the instrumental measurements and mathematical formulations. Most of the physical inversion process do not have a unique solution because of complexity, high-dimensionality, discontinuity, sparsity in the model; or measurement error in the observed data. Therefore, instead of trying to find the exact solution of the inverse problem, it can be easier to understand its characteristics from a distribution. This distribution can be described comprehensively with the posterior probability density function (pdf) of model which contains the prior knowledge and the likelihood which is the misfit between observed and predicted data [3].

Bayes' theorem describes the posterior probability [4]. Calculating the posterior via Bayes' theorem is simple, but it requires sampling from the model space which means solving the forward relation for each sample. Grid search is also possible, however it is not practical for a high-dimensional inverse problem. To decrease the computational cost and time, we need an effective sampling instead of grid search to estimate the model with less samples [5]. Bayesian inference methods are increasingly preferred for solving geophysical problems, thanks to the developments of computational era and effective sampling strategies [6,7]. One can use several various algorithms for sampling which is derivative-free or not. For example, neighborhood algorithm [8,9] and genetic algorithm [10] are derivative-free. However, it is not guaranteed to find the optimal model for global scales.

Hamiltonian Monte Carlo (HMC) algorithm can work well in a global scale, high-dimensional problem, however it requires calculating derivative [11, 12]. Its efficiency is more apparent in the case that derivatives can be calculated rapidly. Surely, there

is no single best optimization algorithm according to the *No Free Lunch theorem* [13]. If taking derivative of the forward relation is easy, then we can say that HMC can be used as an effective sampling method [5].

One of the advantages of describing Hamiltonian dynamics using symplectic geometry is that all of the properties and theorems can be stated geometrically [14]. In view of geometrical perspective, it can be better understood why computational algorithms work well and how we can improve significantly. The characteristics of Hamiltonian dynamics is that the problem is solved in a higher dimensional space (phase space) with additional the generalized momentum $\boldsymbol{p}$ to the generalized position $\boldsymbol{q}$. Hence, the problem can be solved in a more abstract, naive and easy way. The phase space consists of the pair of numbers $(\boldsymbol{q}, \boldsymbol{p})$. As an example, if the physical phenomena occur in the three-dimensional space, e.g. wave propagation then the computations are done in the six-dimensional space. The advantages of transforming the problem to the phase space are based on several geometric results. Firstly, Hamiltonian value is invariant in phase space due to the existence of a symmetry in Hamiltonian equations. Secondly, it is possible to find the Hamiltonian vector field along which Hamiltonian function is invariant. Furthermore, Liouville theorem states that volumes are also preserved along the integral curves of the Hamiltonian vector field. Hence, this abstraction provides us tools to solve not only mechanical problems but also optimization problems such as Hamiltonian Monte Carlo method.

In this thesis, we first describe Hamiltonian dynamics by using symplectic geometry in Section 2.1. We also describe Lagrangian mechanics and use Legendre transformation to obtain Hamiltonian function and Hamilton's equations of motion in Section 2.2. Then, in Chapter 3, we explain how to use Hamiltonian dynamics in a probabilistic inversion problem. More precisely, we define probabilistic spaces, Markov chains and state Ergodic theorem which enables us to draw sample from a distribution and to evaluate the approximate expectations by using samples. The method of Hamiltonian Monte Carlo is explained comprehensively in Chapter 4, and then in Chapter 5, we implement HMC algorithm to tomography problem. Lastly, we explain the use of

HMC in a tomography problem by considering the associated mathematical spaces. More precisely, the physical ray-tracing problem considered is a two dimensional-space representing a subsurface; the dimension of the model space (configuration space) is the number of grids in the physical space and also the phase space. The tomography problem using HMC works in all these spaces with different mathematical structures.

2. HAMILTONIAN DYNAMICS AND SYMPLECTIC GEOMETRY

Hamiltonian dynamics originated as an advanced and relatively abstract formulation of classical mechanics which describes the motion of a system. It can be shown that it is totally equivalent to the Newtonian mechanics. The main difference is that Hamiltonian dynamics uses the energy of the system to describe its motion, whereas the Newtonian mechanics uses the force. Hamiltonian dynamics introduces a new geometric perspective on the mechanics of a system by describing the problem in an extensive space rather that the configuration space of the system, in which a single point represents a state of the whole system at a specific time. In this thesis, the terms configuration space, state space and model space will be used as synonymous in different mathematical contents. The extensive space is called phase space (cotangent bundle) where the coordinates are generalized position and generalized momentum. In order to use the Hamiltonian dynamics for a system, it is necessary to find a scalar function which is constant during the evolution of the system. This function may vary according to the physical phenomena. For example, Hamilton function is total energy for classical mechanics, eikonal equation for ray tracing, and canonical distribution for Hamiltonian Monte Carlo method, as summarized in the next table.

Table 2.1. Hamiltonian for Different Phenomena.

	Classical Mechanics	**Simple Harmonic Oscillation**	**Ray Tracing**	**HMC for Tomography**
$\mathbf{q}$(t)	position of the particle at time t	position of the particle at time t	position of the ray at time t	$\boldsymbol{m}$: velocity model
$\mathbf{p}$ Dual of $\dot{\boldsymbol{q}}$	$\boldsymbol{p} = \frac{\partial K}{\partial \dot{\boldsymbol{q}}}$	$p = m\dot{x}$	wavefront normal $\boldsymbol{p} = \nabla\psi(\boldsymbol{q})$	$\boldsymbol{p} = M\dot{\boldsymbol{m}}$
Constant Fnc.	Total Energy	Total Energy	Eikonal	Canonical Dist.
$H(\boldsymbol{q}, \boldsymbol{p})$	$U(\boldsymbol{q}) + K(\boldsymbol{p})$	$\frac{p^2}{2m} + \frac{kx^2}{2}$	$p^2 - \frac{1}{v^2(\boldsymbol{q},\boldsymbol{p})}$	$\frac{1}{2}\boldsymbol{p}^T M^{-1}\boldsymbol{p} - log\rho_m(\boldsymbol{m}\|\boldsymbol{d})$
Separability into $\boldsymbol{q}$ and $\boldsymbol{p}$	YES	YES	NO	YES
Dimension	1D, 2D, or 3D	1D	2D or 3D	Number of grids

Hamiltonian function is defined on the cotangent bundle of the configuration space (which is also called phase space). The cotangent bundle can be defined as the set of all cotangent spaces at all point $\boldsymbol{q}$ in the configuration space ($\mathbb{R}^n$ for simplicity). The definition of the cotangent space can be given as follows:

Definition 2.0.1. For every $\boldsymbol{q} \in \mathbb{R}^n$, the set of all dual of the tangent vectors $\dot{\boldsymbol{q}}$ at point $\boldsymbol{q}$, denoted by $T^*_{\boldsymbol{q}}\mathbb{R}^n$, is called the *cotangent space* of $\mathbb{R}^n$ at $\boldsymbol{q}$,

$$T^*_{\boldsymbol{q}}\mathbb{R}^n = \{\boldsymbol{p} : T_{\boldsymbol{q}}\mathbb{R}^n \to \mathbb{R} |\ \boldsymbol{p} \text{ is a linear transformation}\}.$$

In other words, $\boldsymbol{p}$ corresponds to projections of the tangent space. That is all linear transformations from tangent space to $\mathbb{R}$ are projection functions. The elements of $T^*_{\boldsymbol{q}}\mathbb{R}^n$ can be called *dual vector*, *one-form*, or *covector*. In Hamiltonian dynamics, it is commonly called *generalized momentum*.

Definition 2.0.2. Let $\boldsymbol{q} \in \mathbb{R}^n$. A *tangent vector* at $\boldsymbol{q}$ is an ordered n-tuple of real numbers $\dot{\boldsymbol{q}} = \langle \dot{q}_1, ..., \dot{q}_n \rangle_{\boldsymbol{q}}$ such that there exists a smooth parameterized curve

$$c : \mathbb{R} \to \mathbb{R}^n$$
$$c(t) = (q_1 + \dot{q}_1 t, ..., q_n + \dot{q}_n t),$$

having the properties that $c(0) = \boldsymbol{q}$ and that

$$c'(0) = \dot{\boldsymbol{q}} = \langle \dot{q}_1, ..., \dot{q}_n \rangle_{\boldsymbol{q}}.$$

Definition 2.0.3. For every $\boldsymbol{q} \in \mathbb{R}^n$, the set of all tangent vectors at $\boldsymbol{q}$ constitutes a vector space. This vector space is called *tangent space* to $\mathbb{R}^n$ at $\boldsymbol{q}$ and denoted by $T_{\boldsymbol{q}}\mathbb{R}^n$.

Definition 2.0.4. The *cotangent bundle* of $\mathbb{R}^n$, denoted by $T^*\mathbb{R}^n$, is the set of all ordered pairs of the form $(\boldsymbol{q}, \boldsymbol{p})$, where $\boldsymbol{q} \in \mathbb{R}^n$ and $\boldsymbol{p} \in T^*_{\boldsymbol{q}}\mathbb{R}^n$.

$$T^*\mathbb{R}^n = \bigcup_{\boldsymbol{q}\in\mathbb{R}^n} T^*_{\boldsymbol{q}}\mathbb{R}^n.$$

It is commonly called *phase space*.

Definition 2.0.5. The *tangent bundle* of $\mathbb{R}^n$, denoted by $T\mathbb{R}^n$, is the set of all ordered pairs of the form $(\boldsymbol{q}, \dot{\boldsymbol{q}})$, where $\boldsymbol{q} \in \mathbb{R}^n$ and $\dot{\boldsymbol{q}} \in T_{\boldsymbol{q}}\mathbb{R}^n$.

$$T\mathbb{R}^n = \bigcup_{\boldsymbol{q}\in\mathbb{R}^n} T_{\boldsymbol{q}}\mathbb{R}^n.$$

A single point $(\boldsymbol{q}, \boldsymbol{p})$ in the phase space contains the generalized position $\boldsymbol{q}$ from the configuration space and the generalized momentum $\boldsymbol{p}$. The generalized position $\boldsymbol{q}$ corresponds to the variable of interest in a system, e.g. it represents the position of the ray path at a specific time in ray tracing problem, while it corresponds to the model parameters in Hamiltonian Monte Carlo Algorithm. Furthermore, generalized momentum $\boldsymbol{p}$ can be defined as any covector in the cotangent space of the configuration space. It does not have to be defined as the multiplication of mass and velocity, that will be be explained later.

2.1. Hamiltonian Equations via Symplectic Form

Hamiltonian dynamics is the main motivation for symplectic geometry. The common point of both approaches is the concept of phase space. As it is mentioned in the previous section, Hamiltonian function is defined on the phase space, i.e., $H : T^*\mathbb{R}^n \to \mathbb{R}$. Furthermore, phase space is the natural setting of symplectic geometry, because we can define a differential two-form on the phase space.

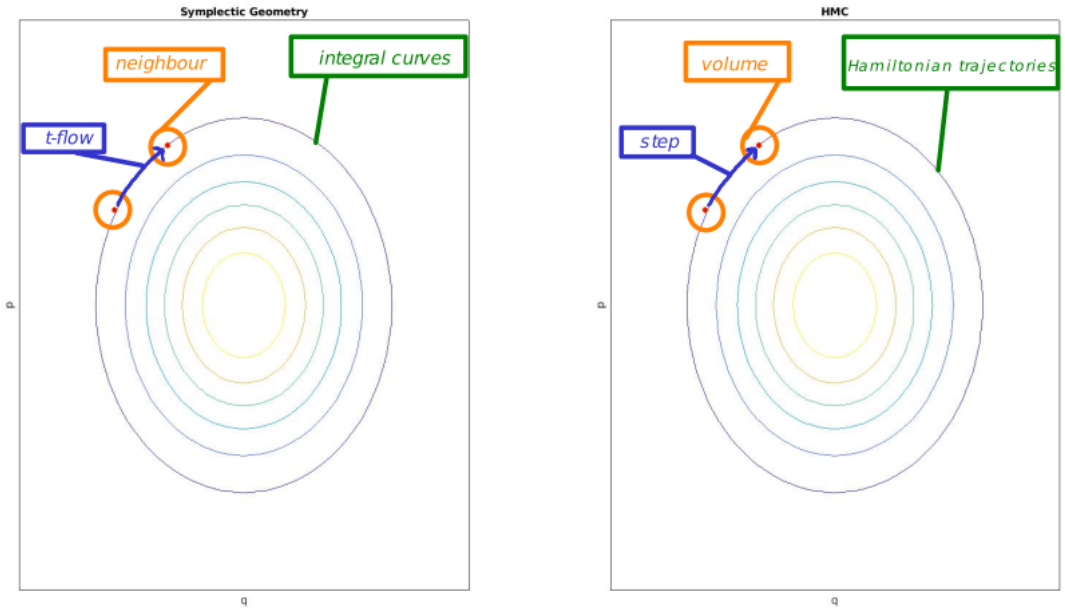

Figure 2.1. The proof of the properties of the HMC method is based on symplectic geometry.

The importance of the Hamiltonian dynamics (accordingly HMC algorithm) is based on its relation with the symplectic geometry as illustrated in Figure 2.1. For that reason, some important definitions about symplectic geometry is going to be given in this section, and their relation with the Hamiltonian dynamics is going to be explained.

Definition 2.1.1. A *symplectic form* on a domain $D \subset T^*\mathbb{R}^n$ is a smooth differential two-form $\boldsymbol{\omega}$ satisfying the following properties:

- $\boldsymbol{\omega}$ is nondegenerate: If $\mathbf{v}_{(\boldsymbol{q},\boldsymbol{p})} \in T_{(\boldsymbol{q},\boldsymbol{p})}D$ has the property that $\boldsymbol{\omega}(\mathbf{v}_{(\boldsymbol{q},\boldsymbol{p})}, \mathbf{w}_{(\boldsymbol{q},\boldsymbol{p})}) = 0$ for all $\mathbf{w}_{(\boldsymbol{q},\boldsymbol{p})} \in T_{(\boldsymbol{q},\boldsymbol{p})}D$, then $\mathbf{v}_{(\boldsymbol{q},\boldsymbol{p})} = \mathbf{0}_{(\boldsymbol{q},\boldsymbol{p})}$.

- $\boldsymbol{\omega}$ is closed: $d\boldsymbol{\omega} = 0$.

Since the elements of the phase space $T^*\mathbb{R}^n$ are taken as a form of $(\boldsymbol{q}, \boldsymbol{p})$, *the standard symplectic form* $\boldsymbol{\omega}$ is considered as

$$\boldsymbol{\omega} = dp_i \wedge dq^i = dp_i \otimes dq^i - dq^i \otimes dp_i, \tag{2.1}$$

for $i \in \{1, ..., n\}$. (Note that the Einstein summation convention is implied for i). Furthermore, the pair $(D, \boldsymbol{\omega})$ is called a *symplectic space*.

Certainly, these requirements for being a symplectic form have mathematical and physical meanings. As a consequence of nondegeneracy condition, we have an isomorphism between tangent bundle TD and the cotangent bundle T^*D. In other words, one can always find a vector field $\mathbf{X}_H$ for any smooth Hamiltonian function H. This isomorphism is explained in following proposition [15],

Proposition 2.1. Let $\mathcal{X}(D)$ be the vector spaces of smooth vector fields on a symplectic space $(D, \boldsymbol{\omega})$, where $D \subset T^*\mathbb{R}^n$ is a domain, and let $\Lambda_1(D)$ be the vector space of one-forms on D. Then the map $\Phi : \mathcal{X}(D) \to \Lambda_1(D)$ given by $\Phi(\mathbf{X}_H) = i(\mathbf{X}_H)\boldsymbol{\omega} = \boldsymbol{\omega}(\mathbf{X}_H, \cdot)$ is a vector space isomorphism.

On the other hand, any vector field is not sufficient to obtain the Hamiltonian equations of motion geometrically. We need a specific vector field along which the symplectic structure $\boldsymbol{\omega}$ is conserved, namely *symplectic vector field* $\mathbf{X}_H$. Mathematically, the Lie derivative of $\boldsymbol{\omega}$ with respect to $\mathbf{X}_H$ should be zero, i.e., $\mathcal{L}_{\mathbf{X}_H}\boldsymbol{\omega} = 0$. Physically, this means that the laws of physics should be independent of time, so that the dynamics is constant along the integral curve of $\mathbf{X}_H$. This can be achieved by the closeness property of the symplectic structure $\boldsymbol{\omega}$ and by the next proposition which is used to define the symplectic vector field $\mathbf{X}_H$.

Proposition 2.2. Let $\mathbf{X}_H$ be a vector field on the symplectic space $(T^*\mathbb{R}^n, \boldsymbol{\omega})$. Then $\mathbf{X}_H$ is a *symplectic vector field* if and only if for each point $(\boldsymbol{q}, \boldsymbol{p}) \in T^*\mathbb{R}^n$, there exists a domain $D \subset T^*\mathbb{R}^n$ containing $(\boldsymbol{q}, \boldsymbol{p})$ and a smooth function $H : D \to \mathbb{R}$ such that

$$i(\mathbf{X}_H)\boldsymbol{\omega} = \boldsymbol{\omega}(\mathbf{X}_H, \cdot) = -dH \quad \text{on D.}$$

Theorem 2.3. (***Liouville Theorem***) Symplectic structure $\boldsymbol{\omega}$ is preserved along the symplectic vector field $\boldsymbol{X}_H$,

$$\mathcal{L}_{\boldsymbol{X}_H}\boldsymbol{\omega} = 0. \tag{2.2}$$

Hence, volume-form $\boldsymbol{\omega}^n$ in phase space is preserved along the symplectic vector field $\mathbf{X}_H$, i.e.,

$$\mathcal{L}_{\boldsymbol{X}_H}\boldsymbol{\omega}^n = 0. \tag{2.3}$$

Lie Derivative (Cartan's Formula)

The Lie derivative can be expressed by the Cartan's formula as following,

$$\mathcal{L}_{\mathbf{X}_H}\boldsymbol{\omega} = i(\mathbf{X}_H)d\boldsymbol{\omega} + d(i(\mathbf{X}_H)\boldsymbol{\omega}).$$

Proof. By Proposition 2.1 and Proposition 2.2,

$$\mathcal{L}_{\mathbf{X}_H}\boldsymbol{\omega} = i(\mathbf{X}_H)d\boldsymbol{\omega} + d(-dH). \tag{2.4}$$

The second term is vanished, since $d^2 = 0$

$$\mathcal{L}_{\mathbf{X}_H}\boldsymbol{\omega} = i(\mathbf{X}_H)d\boldsymbol{\omega}. \tag{2.5}$$

Evidently, the fact that the Lie derivative is zero depends on the fact that the symplectic structure is closed, i.e, $d\boldsymbol{\omega} = 0$ and on the way to define the symplectic vector field $\mathbf{X}_H$. By using this fact, the Lie derivative of the volume form can be rewritten by product rule,

$$\mathcal{L}_{\mathbf{X}_H}\boldsymbol{\omega}^n = \mathcal{L}_{\mathbf{X}_H}\boldsymbol{\omega}^{n-1} \wedge \boldsymbol{\omega} + \boldsymbol{\omega}^{n-1} \wedge \mathcal{L}_{\mathbf{X}_H}\boldsymbol{\omega}. \tag{2.6}$$

The second term is vanished by Equation (2.2). By applying product rule again,

$$\begin{aligned}
\mathcal{L}_{\mathbf{X}_H}\boldsymbol{\omega}^n &= \mathcal{L}_{\mathbf{X}_H}\boldsymbol{\omega}^{n-1} \wedge \boldsymbol{\omega} \\
&= \mathcal{L}_{\mathbf{X}_H}\boldsymbol{\omega}^{n-2} \wedge \boldsymbol{\omega} \\
&= \mathcal{L}_{\mathbf{X}_H}\boldsymbol{\omega}^{n-3} \wedge \boldsymbol{\omega} \\
&\vdots \\
&= \mathcal{L}_{\mathbf{X}_H}\boldsymbol{\omega} \wedge \boldsymbol{\omega} \\
&= 0.
\end{aligned} \tag{2.7}$$

□

Liouville Theorem 2.3 is used for computational part; discretized probability function is approximated as volume times probability function. It provides a great convenience, since it states that the symplectic form is invariant along the vector field, i.e., a sufficiently small volume on the vector field is preserved along the integral curves of the vector field. In other words, it is not necessary to consider the volume during discretizing the probability function, since the volume is equal for each state.

By using the definitions and propositions which is stated, we can obtain Hamiltonian equations of motion from any smooth Hamiltonian function. Suppose the Hamiltonian function $H\colon T^*\mathbb{R}^n \to \mathbb{R}$ is given by $(\boldsymbol{q}, \boldsymbol{p}) \mapsto \mathbb{R}$, then its differential is

$$dH = \frac{\partial H}{\partial q^i} dq^i + \frac{\partial H}{\partial p_i} dp_i. \tag{2.8}$$

Firstly, define the symplectic vector field $\mathbf{X}_H$ by using Proposition 2.2

$$-dH = \boldsymbol{\omega}(\mathbf{X}_H, \cdot) = dp_i \wedge dq^i(\mathbf{X}_H, \cdot). \tag{2.9}$$

By using Equation (2.1),

$$-dH = dp_i(\mathbf{X}_H) \otimes dq^i - dq^i(\mathbf{X}_H) \otimes dp_i. \tag{2.10}$$

Since $\mathbf{X}_H$ is a vector field on $T(T^*\mathbb{R}^n)$, a vector on $\mathbf{X}_H$ is taken form of $(\dot{\boldsymbol{q}}, \dot{\boldsymbol{p}})$

$$\begin{aligned} -dH &= dp_i((\dot{\boldsymbol{q}}, \dot{\boldsymbol{p}})) \otimes dq^i - dq^i((\dot{\boldsymbol{q}}, \dot{\boldsymbol{p}})) \otimes dp_i \\ &= \dot{p}_i dq^i - \dot{q}^i dp_i = -\frac{\partial H}{\partial q^i} dq^i - \frac{\partial H}{\partial p_i} dp_i, \end{aligned} \tag{2.11}$$

where $\boldsymbol{\omega}$ is standard symplectic form. Hence, we obtain the Hamiltonian equation of motion as follows,

$$\frac{\partial H}{\partial \boldsymbol{q}} = -\dot{\boldsymbol{p}} \qquad\qquad \frac{\partial H}{\partial \boldsymbol{p}} = \dot{\boldsymbol{q}} \tag{2.12}$$

Obviously, there is a symmetry between $\boldsymbol{q}$ and $\boldsymbol{p}$. Noether's theorem states that existence of a symmetry implies existence of a corresponding invariant of the system [16], so there is an invariant quantity in the Hamiltonian dynamics. Each invariant reduces one number of degree of freedom of the system. Hamiltonian $H(\boldsymbol{q}, \boldsymbol{p})$ is the corresponding invariant of the Hamiltonian dynamics and it makes possible to solve the problem along the trajectories by fixing momentum, where the Hamiltonian is constant in the

phase space.

Definition 2.1.2. Let $(\boldsymbol{V}_1, \boldsymbol{\omega}_1)$ and $(\boldsymbol{V}_2, \boldsymbol{\omega}_2)$ be two symplectic spaces, for domains $\boldsymbol{V}_1, \boldsymbol{V}_2 \subset T^*\mathbb{R}^n$, and let $\Phi : \boldsymbol{V}_1 \to \boldsymbol{V}_2$ be a diffeomorphism. Then Φ is called a *symplectic diffeomorphism* if

$$\Phi^* \boldsymbol{\omega}_2 = \boldsymbol{\omega}_1.$$

We write $\Phi : (\boldsymbol{V}_1, \boldsymbol{\omega}_1) \to (\boldsymbol{V}_2, \boldsymbol{\omega}_2)$.

Those diffeomorphisms have a great importance as they preserve the symplectic structure, hence the volume. Darboux's theorem guarantees that such an isomorphism Φ in Definition 2.1.2 always exists.

Darboux's Theorem for the symplectic geometry

Theorem 2.4. Let $(T^*\mathbb{R}^n, \boldsymbol{\omega})$ be a symplectic space. For each $(\boldsymbol{q}, \boldsymbol{p}) \in T^*\mathbb{R}^n$, there exists a domain D containing $(\boldsymbol{q}, \boldsymbol{p})$ and a diffeomorphism $\Phi : D \to \Phi(D) \subset T^*\mathbb{R}^n$ such that $\Phi((\boldsymbol{q}, \boldsymbol{p})) = (\mathbf{q}, \mathbf{p})$ and on D, $\Phi^*\boldsymbol{\omega} = \boldsymbol{\omega}$, where $\boldsymbol{\omega} = dp_i \wedge dq^i$ is the standard symplectic form on $T^*\mathbb{R}^n$ with coordinate pairs (q^i, p_i).

More intuitively, the Darboux's theorem guarantees the existence of the diffeomorphism between any two symplectic structures in the phase space, which is a great convenience for using the standard symplectic form for the phase space.

In order to be more intuitive and less abstract, it is convenient to acquire the Hamiltonian equations for classical mechanics by using Lagrangian dynamics, since it is the physical phenomena that everyone is most accustomed. The next section is written for this reason.

2.2. Hamiltonian Equations via Lagrangian

The aim of this chapter is to show the equality of Newtonian dynamics and the Hamiltonian dynamics in such a less abstract way, by changing variables of Newtonian dynamics to find the Hamiltonian equations.

In Newtonian dynamics, the motion of a particle can be described by a second-order differential equation,

$$\boldsymbol{F}(\boldsymbol{q}, \dot{\boldsymbol{q}}, t) = m\ddot{\boldsymbol{q}}, \tag{2.13}$$

where m is mass and $\boldsymbol{q}$, $\dot{\boldsymbol{q}}$, and $\ddot{\boldsymbol{q}}$ are the position, velocity, and acceleration of the particle, respectively. In order to achieve equation of motion for a physical phenomena, one has to consider a great number of vectors (forces) acting on the particle, i.e. determining their directions and norms.

Alternatively, it is possible to transform from Newtonian dynamics to Hamiltonian dynamics by changing variables. There is an intermediate stage, namely Lagrangian dynamics, during this transition. The configuration space $\mathbb{R}^n$ consists of the points that represent the whole situation of the particles or rigid bodies in Newtonian dynamics. As time evolves, the position of the rigid bodies or particles changes and the point $\boldsymbol{q} \in \mathbb{R}^n$ in the configuration space moves along a curve. The Lagrangian function $L(\boldsymbol{q}, \dot{\boldsymbol{q}})$ is defined on $T\mathbb{R}^n$, the tangent bundle of the configuration space. Though $\boldsymbol{q}$ and its time derivative $\dot{\boldsymbol{q}}$ are functionally dependent, $\boldsymbol{q}$ and $\dot{\boldsymbol{q}}$ treat as independent of each other, since $\dot{\boldsymbol{q}}$ can be any vector in the tangent space of $\boldsymbol{q}$, $T_{\boldsymbol{q}}\mathbb{R}^n$ [15].

Theorem 2.5. For position $\boldsymbol{q} \in \mathbb{R}^n$, any vector $\dot{\boldsymbol{q}} = \begin{bmatrix} \dot{q}_1 & \dot{q}_2 & \cdots & \dot{q}_n \end{bmatrix}^T$ can be regarded as a tangent vector at $\boldsymbol{q}$.

Proof. For $\boldsymbol{q} = (q_1, q_2, ..., q_n)$, define a curve $c : \mathbb{R} \to \mathbb{R}^n$ such that

$$c(t) = (q_1 + \dot{q}_1 t, ..., q_n + \dot{q}_n t).$$

Then, it is obvious that $c(0) = \boldsymbol{q}$ and $c'(0) = \dot{\boldsymbol{q}}$. Hence, $\dot{\boldsymbol{q}} \in T_{\boldsymbol{q}}\mathbb{R}^n$. □

In classical mechanics, the Lagrangian function is defined as the difference of the kinetic and the potential energies to get the equation of the motion, because of the fact that it satisfies the least action principle.

$$\begin{aligned} L : T\mathbb{R}^n &\to \mathbb{R} \\ L(\boldsymbol{q}, \dot{\boldsymbol{q}}) &= K(\dot{\boldsymbol{q}}) - U(\boldsymbol{q}), \end{aligned} \tag{2.14}$$

where K is the kinetic energy of the system and U is the potential energy of the system. This simple scalar equation can summarize the dynamics of the entire system.

$$\begin{aligned} K &= \frac{1}{2} m \dot{\boldsymbol{q}}^2 \\ U &= U(\boldsymbol{q}) \\ L &= K - U = \frac{m}{2} \dot{\boldsymbol{q}}^2 - U(\boldsymbol{q}) \end{aligned}$$

The goal is to find the position $\boldsymbol{q}$, which is the solution of the equations of motion, hence it should satisfy the least action principle, i.e. $\boldsymbol{q}$ should yield a stationary value of the integral of the Lagrangian over time, namely action, S.

$$S[\boldsymbol{q}] = \int_{t_1}^{t_2} L(\boldsymbol{q}, \dot{\boldsymbol{q}}) dt \tag{2.15}$$

How does the least action principle imply the Euler-Lagrange equations?
Suppose that $\boldsymbol{q}$ is a stationary point of S, and that it is parameterized on $[t_1, t_2] \subset \mathbb{R}$ as

$$\boldsymbol{q}(t;\varepsilon) = \boldsymbol{q}(t) + \varepsilon f(t)$$

where ε is just a number and $f(t)$ is any arbitrary function which satisfies $f(t_1) = f(t_2) = 0$ (so that it does not change the boundary values). One can easily recognize that $\varepsilon = 0$ corresponds to stationary path $\boldsymbol{q}(t)$, i.e.

$$\left.\frac{\partial S}{\partial \varepsilon}\right|_{\varepsilon=0} = 0$$

Let's take derivative of the action S with respect to ε in Equation (2.15),

$$\begin{aligned}
\frac{\partial S}{\partial \varepsilon} &= \int_{t_1}^{t_2} \frac{\partial L}{\partial \boldsymbol{q}} \frac{d\boldsymbol{q}}{d\varepsilon} + \frac{\partial L}{\partial \dot{\boldsymbol{q}}} \frac{d\dot{\boldsymbol{q}}}{d\varepsilon} dt \\
&= \int_{t_1}^{t_2} \frac{\partial L}{\partial \boldsymbol{q}} f(t) + \frac{\partial L}{\partial \dot{\boldsymbol{q}}} \dot{f}(t) dt \\
&= \int_{t_1}^{t_2} \frac{\partial L}{\partial \boldsymbol{q}} f(t) dt + \left.\frac{\partial L}{\partial \dot{\boldsymbol{q}}} f(t)\right|_{t_1}^{t_2} - \int_{t_1}^{t_2} f(t) \frac{d}{dt} \frac{\partial L}{\partial \dot{\boldsymbol{q}}} dt \\
&= \left.\frac{\partial L}{\partial \dot{\boldsymbol{q}}} f(t)\right|_{t_1}^{t_2} + \int_{t_1}^{t_2} \left(\frac{\partial L}{\partial \boldsymbol{q}} - \frac{d}{dt} \frac{\partial L}{\partial \dot{\boldsymbol{q}}} \right) f(t) dt \\
&= 0
\end{aligned}$$

The first term vanishes because $f(t)$ vanishes at the endpoints. Since $f(t)$ is an arbitrary function, the only way for the second term to vanish is to satisfy the Euler-Lagrange equations, i.e.

$$\frac{d}{dt} \frac{\partial L}{\partial \dot{\boldsymbol{q}}} = \frac{\partial L}{\partial \boldsymbol{q}}$$

Theorem 2.6. A smooth function $\boldsymbol{q} : I \to \mathbb{R}^n$ described as $\boldsymbol{q}(t) = (q_1(t), ..., q_2(t))$ is a solution to the equations of motion described by the system $\ddot{\boldsymbol{q}} = -\nabla U$ with potential energy U, kinetic energy T, and the Lagrangian function $L(\boldsymbol{q}, \dot{\boldsymbol{q}}) = T(\dot{\boldsymbol{q}}) - U(\boldsymbol{q})$, if and only if

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = 0, \ i = 1, ...n, \tag{2.16}$$

where the n equations are known as the *Euler-Lagrange equations* [15].

Lagrangian equation of motion is more abstract and practical than Newtonian dynamics, since one has to concern with one scalar function instead of considering several force vectors.

However, one still has to solve a second order differential equations system, in Equation (2.16). There exists a more useful and naive way to describe the equation of motion, namely Hamiltonian dynamics. Unlike Lagrangian, Hamiltonian is invariant, and it reduces the degree of freedom. It allows us to solve the problem along the Hamiltonian trajectories.

Hamiltonian equations can be inferred from Lagrangian dynamics. While the variable of Lagrangian is position $\boldsymbol{q}$ and velocity $\dot{\boldsymbol{q}}$, Hamiltonian is a function of position $\boldsymbol{q}$ and momentum $\boldsymbol{p}$.

Legendre transformation is used to convert Lagrangian function $L(\boldsymbol{q}, \dot{\boldsymbol{q}})$, which is defined on the tangent space $T_{\boldsymbol{q}}\mathbb{R}^n$, into the Hamiltonian function $H(\boldsymbol{q}, \boldsymbol{p})$, which is defined on the cotangent space $T^*_{\boldsymbol{q}}\mathbb{R}^n$ (the dual of the tangent space). The interchanging variables $(\dot{\boldsymbol{q}}, \boldsymbol{p})$ are conjugate pair of variables, where $\boldsymbol{p}$ is the momentum of the system. Since the transformation preserves the unit of the function, both of the Hamiltonian and Lagrangian have the unit of energy. Legendre transform from

Lagrangian to Hamiltonian is simply formulated as

$$H(\boldsymbol{q}, \boldsymbol{p}) = \boldsymbol{p}\dot{\boldsymbol{q}} - L(\boldsymbol{q}, \dot{\boldsymbol{q}}). \tag{2.17}$$

Here, the necessary condition is that the Hamiltonian function should be independent of $\dot{\boldsymbol{q}}$. In other words, when its differential is considered, the coefficient of $d\dot{\boldsymbol{q}}$ should be vanished,

$$\begin{aligned} dH &= \boldsymbol{p}d\dot{\boldsymbol{q}} + \dot{\boldsymbol{q}}d\boldsymbol{p} - \frac{\partial L}{\partial \boldsymbol{q}}d\boldsymbol{q} - \frac{\partial L}{\partial \dot{\boldsymbol{q}}}d\dot{\boldsymbol{q}} \\ &= \dot{\boldsymbol{q}}d\boldsymbol{p} + \left(\boldsymbol{p} - \frac{\partial L}{\partial \dot{\boldsymbol{q}}}\right) d\dot{\boldsymbol{q}} - \frac{\partial L}{\partial \boldsymbol{q}}d\boldsymbol{q} \end{aligned} \tag{2.18}$$

In order to make the coefficient of $d\dot{\boldsymbol{q}}$ zero, the generalized momentum should be defined as

$$\boldsymbol{p} = \frac{\partial L}{\partial \dot{\boldsymbol{q}}} \tag{2.19}$$

Also, from Equation (2.16), we obtain

$$\dot{\boldsymbol{p}} = \frac{d\boldsymbol{p}}{dt} = \frac{d}{dt}\frac{\partial L}{\partial \dot{\boldsymbol{q}}} = \frac{\partial L}{\partial \boldsymbol{q}} \tag{2.20}$$

Then, substituting $\dot{\boldsymbol{p}}$ into the Equation (2.18)

$$\begin{aligned} dH &= \dot{\boldsymbol{q}}d\boldsymbol{p} - \dot{\boldsymbol{p}}d\boldsymbol{q} \\ \frac{\partial H}{\partial \boldsymbol{q}}d\boldsymbol{q} + \frac{\partial H}{\partial \boldsymbol{p}}d\boldsymbol{p} &= \dot{\boldsymbol{q}}d\boldsymbol{p} - \dot{\boldsymbol{p}}d\boldsymbol{q} \end{aligned} \tag{2.21}$$

Thus, Hamiltonian equations of motion can be written as a system of first order differential equations as in Equation (2.12),

$$\frac{\partial H}{\partial \boldsymbol{q}} = -\dot{\boldsymbol{p}} \qquad\qquad \frac{\partial H}{\partial \boldsymbol{p}} = \dot{\boldsymbol{q}} \tag{2.22}$$

In summary, we need to solve six first-order differential equations using Hamiltonian, instead of three second-order differential equations for a three dimensional system, which gives us more simple equations. Furthermore, Hamiltonian equations have a symmetry, unlike the Euler-Lagrange equations. Since this reduces the degree of freedom, the problem can be solved along the constant Hamiltonian trajectories.

We can show that the Hamiltonian function corresponds to the total energy of a particular physical system. It can be derived as follows,

$$H(\boldsymbol{q}, \boldsymbol{p}) = \boldsymbol{p}\dot{\boldsymbol{q}} - \left(\frac{m}{2}\dot{\boldsymbol{q}}^2 - U(\boldsymbol{q})\right) \tag{2.23}$$

since $\boldsymbol{p} = \frac{\partial L}{\partial \dot{\boldsymbol{q}}} = m\dot{\boldsymbol{q}}$, substitute $\dot{\boldsymbol{q}}$ into the above equation,

$$\begin{aligned} H(\boldsymbol{q}, \boldsymbol{p}) &= \frac{\boldsymbol{p}^2}{m} - \left(\frac{\cancel{m}}{2}\frac{\boldsymbol{p}^2}{\cancel{m}^2} - U(\boldsymbol{q})\right) \\ &= \frac{\boldsymbol{p}^2}{2m} + U(\boldsymbol{q}) \\ &= K(\boldsymbol{p}) + U(\boldsymbol{q}) \end{aligned} \tag{2.24}$$

Surely, the Hamiltonian does not have to be total energy of the system. In Table 2.1, there are some examples for the Hamiltonian for three different phenomena. For HMC method to solve an inverse problem, we will define the Hamiltonian as the total energy of an artificial Hamiltonian dynamics.

3. HAMILTONIAN MONTE CARLO FOR INVERSION

The process for estimating model parameters from observed data is called *inversion*. Finding an exact solution of an inverse problem is not always possible since there might be some noises in the measured data. On the other hand, to test all possible solutions to find the best solution is not practical. However, we can select sufficient amount of acceptable models, from which we can estimate the probability characteristic of the model parameters. In such a case, the histogram of the all acceptable models corresponds to the density defined over the model space.

In order to solve an inverse problem, firstly one should know how the model parameters are related with the data. For travel-time tomography, the relation between model parameters and data is given by Hamiltonian equations of motion. In other words, Hamiltonian equations of motion relates the velocity structure (model parameters) and the travel-time of seismic waves (data). There are various ways to solve inverse problems, namely in particular the gradient based methods and Monte Carlo methods. In both cases, the best model is chosen by analyzing the error between observed data and the predicted data. Monte Carlo methods is based on sampling model parameters and then evaluating the predicted data of the model via forward relation.

The second step of solving an inverse problem is to find an efficient sampling method to walk in model space. This step requires to tune some parameters such as step-size and number of steps in order to achieve effective sampling. Once the parameters are tuned properly, the inversion process finishes in a reasonable amount of time.

There are several Monte Carlo methods for sampling the model space, such as Markov Chain Monte Carlo (MCMC) and Hamiltonian Monte Carlo (HMC). The main difference of these two methods is that MCMC makes a random walk, whereas HMC walks along a specific vector field, which is so-called Hamiltonian vector field. The

advantage of this geometrical approach is that HMC makes more efficient sampling, i.e. it has a higher acceptance rate than MCMC methods. This higher acceptance of HMC results in less iteration, less computational time.

These two steps, namely determining the forward relation and sampling, are combined with the misfit function, which returns the error between predicted and observed data. Finally, the global minimum of the misfit function is the solution of the inverse problem. However, HMC does not find the global minimum of the misfit function directly, instead it reveals the target probability distribution of the model parameters by using the misfit function. HMC samples according to this distribution so that one can evaluate the expectation of the distribution by using these samples. As an example, if the target distribution is defined as Gaussian, then the best solution can be find by taking the average of samples (expected value of the distribution). These results, namely sampling from a distribution and the average of the samples converges to the expectation, are guaranteed by Ergodic Theorem which will be explained in Section 3.2 in detail.

In order to state Ergodic theorem, we introduce basic definitions and results of probability theory and then define a joint distribution in the phase space in the next section.

3.1. Probability Theory

A *random variable* can be defined intuitively as a variable which is used to represent the outcome of a random experiment (measurement), which can either be a scalar or a vector-valued depending on the model. Therefore, it is possible to think of models and data as random variables. Random variables cannot give useful information about the relevant process individually. However, with a large number of measurements, one can get a distribution to characterize the model parameters completely. It means that one can use probability theory as a tool in order to assign a numerical value to each state during an inversion process, namely the numerical value corresponds to the prob-

ability of that state where the states are the elements of the model space. Hence, in this section we give basic definitions of probability theory.

Firstly, we need to define a *measure* on the data to analyze how likely of an event is to occur after large number of measurements.

Definition 3.1.1. A *probability measure* $(S, \mathcal{A}, P)$ on a set S with $\sigma - algebra$ $\mathcal{A}$ is a function $P : \mathcal{A} \to [0, 1]$ that satisfies Kolmogorov axioms:

- $P(\varnothing) = 0$.
- $P(\bigcup_{i \in \infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$ for any $E_i \in \mathcal{A}$ of pairwise disjoint sets.
- $P(S) = 1$.

where σ-*algebra* $\mathcal{A}$ on S is a nonempty subset of the power set of S which is closed under complements, and closed under countable unions. In HMC algorithm, the set S corresponds to the state space (model space).

Definition 3.1.2. Let X be a random variable. X is said to have a normal distribution, if its probability density function (pdf) is given as

$$\rho(X) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right), \tag{3.1}$$

where $\mu = \mathbb{E}[X] \in \mathbb{R}$ and $\sigma^2 = \mathbb{E}[(X-\mu)^2] \in \mathbb{R}$ are the *mean* and *variance*, respectively. It is denoted by $X \sim \mathcal{N}(\mu, \sigma)$. Also

$$\int_{-\infty}^{\infty} \rho(X) dX = 1. \tag{3.2}$$

The mean of a continuous random variable X is defined as

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} X\rho(X) dX. \tag{3.3}$$

By Equations (3.2) and (3.3), the mean (expected value) of a random variable can be defined as the weighted average of all possible outcomes. It is possible to generalize this normal distribution to higher dimensions, namely multivariate normal distribution

Definition 3.1.3. Let $\boldsymbol{X} = [X_1, X_2, ..., X_n]^T$ be a n-dimensional random variable. $\boldsymbol{X}$ is said to be have a normal distribution, $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if its pdf is given as

$$\rho(\boldsymbol{X}) = \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}|}} \exp\left(-\frac{(\boldsymbol{X}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})}{2}\right)$$

where $\boldsymbol{\mu}$ is the n-dimensional mean vector, $\boldsymbol{\Sigma}$ is the $n \times n$ covariance matrix, and the $|\boldsymbol{\Sigma}|$ is the determinant of the covariance matrix,

$$\boldsymbol{\mu} = E[\boldsymbol{X}] = [E[X_1], E[X_2], ..., E[X_n]]^T$$

$$\boldsymbol{\Sigma} = \sum_{i=1}^{n}\sum_{j=1}^{n}\Sigma_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{n}E[(X_i-\mu_i)(X_j-\mu_j)]$$

Remark 3.1.1. If random variables are discrete, then the integration corresponds to the summation, and it is called probability mass function (pmf), rather than pdf.

Probability of a state s, $P(X = s)$ represents the marginal probability of being the state s in a set. Furthermore, the conditional probability $P(X = s \mid Y = d)$ is defined as the probability of being state s when the state d is already occurred in another set, where X and Y are random variables of two different sets.

Theorem 3.1 (**Bayes' theorem**)**.** It states that there is a relation between conditional probabilities $P(X \mid Y)$ and $P(Y \mid X)$ such that

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}.$$

Bayes' theorem is used commonly for inversion methods, where X and Y are corresponds to the random variables of model space and data space, respectively.

Lastly, the joint probability distribution $P(X, Y)$ corresponds the probability the all possible pairs of outcomes of the two sets. In the case that X and Y are independent random variables, then the joint probability of them is multiplication of the marginal probabilities $P(X)$ and $P(Y)$.

3.2. Markov Chain Monte Carlo

The aim of a Markov Chain Monte Carlo (MCMC) is to sample from a given distribution. Hamiltonian Monte Carlo is an MCMC algorithm that does random walk according to the distribution which is based on Hamiltonian dynamics. A Markov chain must have a stationary distribution to ensure that one can sample from that distribution by using Hamiltonian dynamics. The existence of a stationary (invariant) distribution is guaranteed by the detailed balance property. In order to understand what HMC and its advantages are, we start with the definition of Markov chain and necessary properties of Markov Chain Monte Carlo and the hypothesis of Ergodic theorem.

Definition 3.2.1. A sequence $S_1, S_2, \cdots$ of random elements in state space $\mathcal{C}$ is a *Markov Chain* if the conditional (transition) probability satisfies

$$T(S_{i+1} \mid S_i, S_{i-1}, \cdots, S_1) = T(S_{i+1} \mid S_i). \tag{3.4}$$

More intuitively, Markov Chain can be defined as a process of random sampling in which the choice of the current state depends only on the the previous state. Although this property has a great importance for requiring less memory, it is not sufficient for the existence of a stationary distribution. We need a Markov chain that satisfies the Ergodic theorem.

Theorem 3.2 (**Ergodic theorem**). If a Markov Chain S_i in a state space is

- time homogeneous,
- irreducible,
- has a stationary (invariant) distribution ρ,

then

$$\mathbb{E}(f(X)) \xrightarrow[n\to\infty]{} \frac{1}{n}\sum_{i=1}^{n} f(S_i), \quad X \sim \rho \tag{3.5}$$

where f is a real-valued function on the state space. Furthermore, if it is aperiodic, then the transition probability of one state to another state in a sufficiently number of steps converges to the stationary distribution, i.e.

$$T(S_n = s \mid S_0 = s_0) \xrightarrow[n\to\infty]{} \rho(s),$$

where s and s_0 are in the state space. Moreover, a Markov Chain that satisfies these properties is called *ergodic*.

Now, we are going to define the properties of a Markov chain that is stated in the hypothesis of Ergodic Theorem 3.2.

Definition 3.2.2. A Markov Chain $\{S_i \mid i \in \mathbb{N}\}$ is called *time-homogeneous*, if the conditional (transition) probability satisfies

$$T(S_{i+1} = b \mid S_i = a) = T_{ab}, \quad \forall i \in \mathbb{N}, \quad \forall a, b \in \mathcal{C},$$

for some transition matrix T, where T can be represented by a square matrix and its each entry corresponds to transition probability from one state to another in Figure 3.1. In order to guarantee that a walker goes to a state in the state space, the sum of

transition probabilities from a state to all states in state space must be unity, that is

$$\sum_b T_{ab} = 1. \tag{3.6}$$

Intuitively, if a transition matrix is time-homogeneous, then the transition probability is not a function of time, i.e. no matter which step of Markov Chain we are in, the probability of transition from state a to state b remains constant as shown in the next figure.

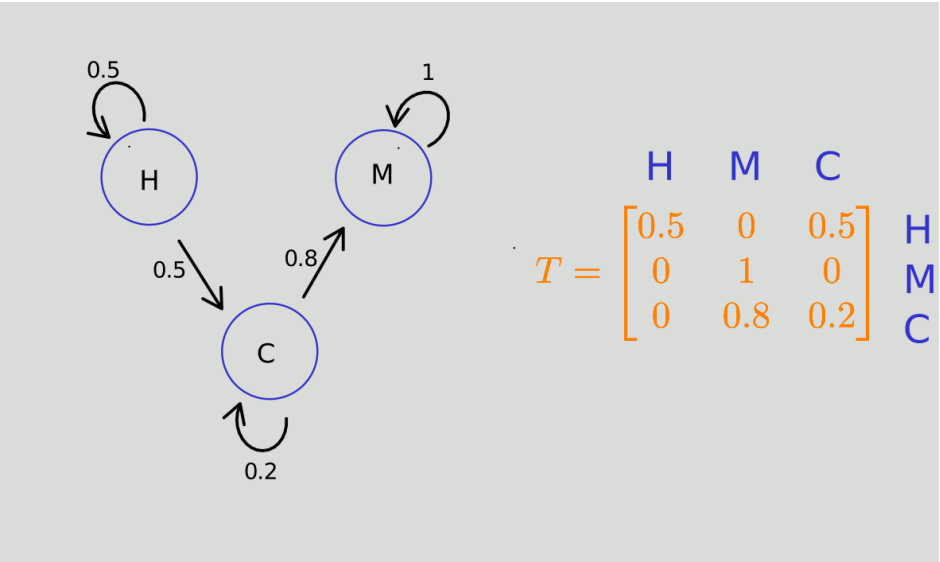

Figure 3.1. The transition probabilities are invariant with respect to time, therefore the transition matrix is time-homogeneous.

Definition 3.2.3. A Markov Chain is called *irreducible* if

$$P(S_t = b \mid S_0 = a) > 0, \quad \exists t \geq 0,$$

for all a and b in the state space.

Definition 3.2.4. An irreducible Markov chain S_i is called *aperiodic*, if

$$gcd\{t : T(S_t = a \mid S_0 = a) > 0\} = 1,$$

for any state a in the state space.

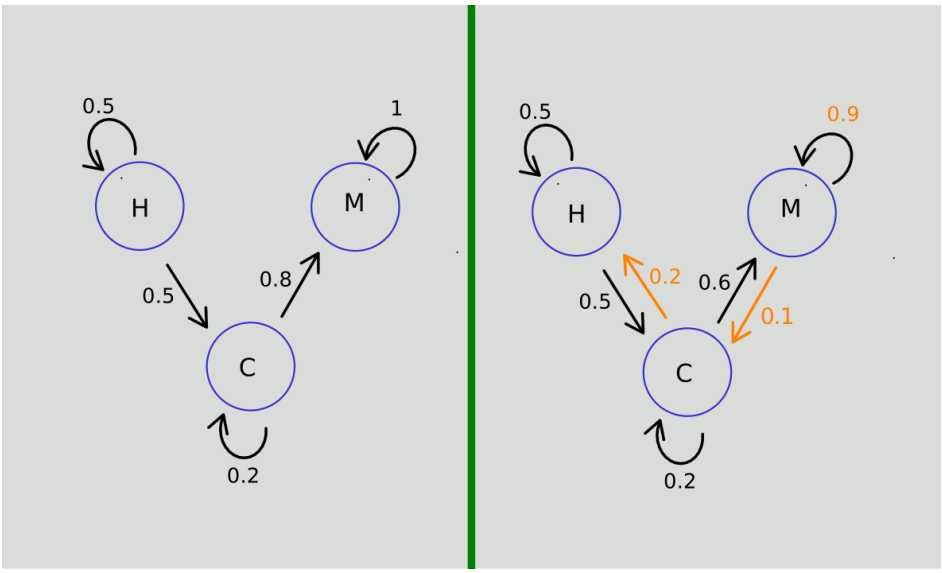

Figure 3.2. Illustration of the irreducibility property.

The properties of irreducibility and aperiodicity imply that we can go to any state from the current state. They prevent us from getting stuck in one state during a random walk in the state space. For example, in Figure 3.2, the Markov chain (MC) on the left side is not irreducible because one cannot go to the state H or C once M is visited. However, the MC on the right side is irreducible: each state can be reached from any other state.

Definition 3.2.5. A probability mass function ρ on the state space $\mathcal{C}$ satisfies *detailed balance* (*reversibility*) with respect to T, if

$$\rho_a T_{ab} = \rho_b T_{ba} \tag{3.7}$$

for all a and b in $\mathcal{C}$.

Definition 3.2.6. A probability mass function ρ is a *stationary distribution* on the state space $\mathcal{C}$ with respect to the transition matrix T if

$$\rho = \rho T$$

In other words, ρ is a stationary distribution, if

$$\rho_b = \sum_a \rho_a T_{ab}$$

Intuitively, ρ is a stationary distribution with respect to T, if it is invariant by the transition matrix T.

Recall that one of the hypothesis of Ergodic Theorem 3.2 is that a Markov chain must have a stationary distribution. It is easier to check detailed balance property which implies the existence of stationary distribution as proved by the following theorem.

Theorem 3.3. Detailed balance ensures the existence of a stationary distribution ρ.

Proof. Suppose that ρ satisfies detailed balance, then by Definition 3.2.5,

$$\sum_a \rho_a T_{ab} = \sum_a \rho_b T_{ba}$$

Since ρ_b is independent of a,

$$\sum_a \rho_a T_{ab} = \sum_a \rho_b T_{ba} = \rho_b \sum_a T_{ba}$$

By Equation (3.6),

$$\sum_a \rho_a T_{ab} = \sum_a \rho_b T_{ba} = \rho_b \sum_a T_{ba} = \rho_b$$

□

3.3. Markov Chain and Sampling

The simplest way to produce a Markov chain for sampling is drawing a new random state from the neighbour of the current state, and making an accept-reject test on the new state depending on the target distribution. For example, the target distribution may be defined as the exponential of misfit function. More generally, once a probability distribution is defined on the model space, then one can use Monte Carlo method to draw random samples from the target distribution. By Theorem 3.2, one can converge to the stationary distribution by using the transition matrix of the Markov chain. It has a crucial importance for analyzing (evaluating the expectation) the distribution of the model parameters in high-dimensional spaces via drawing the large number of samples. However, it converges slowly to the target distribution, as it does not guarantee drawing independent samples, i.e. it needs to collect more samples, to take more steps to walk all over space, to make more accept-reject test, and to increase the rate of accept-reject, etc. These challenges increase the computational time and costs, and may even make solving high dimensional problems impossible.

Hamiltonian Monte Carlo method puts a restriction on sampling, i.e. sampling only on the same Hamiltonian trajectory with small numeric error, with an Hamiltonian function which should be remain constant during walking in the phase space. It makes more effective sampling thanks to this simple but impressive restriction, so that it proposes more distinct samples, explores the space faster, increases the rate of accept-reject to converge to target distribution rapidly.

3.4. Ergodic Markov Chain and HMC Sampling

Hamiltonian Equations (2.12) can be used for sampling as an MCMC thanks to its three crucial properties, namely reversibility, preserving volume, and invariant Hamiltonian. The MCMC which is based on Hamiltonian dynamics is called Hamiltonian Monte Carlo. Accepted samples during HMC sampling form an ergodic Markov chain thanks to these three properties of Hamiltonian dynamics.

Table 3.1. Relations of the properties of three concepts

Phase Space	**Hamiltonian Dynamics**	**HMC**
Diffeomorophism generated by $\boldsymbol{X}_H$	Reversibility	Detailed balance
$\mathcal{L}_{\boldsymbol{X}_H}\boldsymbol{\omega} = 0$	Preserving volume	invariant discrete joint pdf
$\frac{dH}{dt} = 0$	Invariant Hamiltonian	invariant continuous joint pdf

Firstly, Hamiltonian dynamics is reversible, it means that trajectory mapping which is obtained by HMC has an inverse (Figure 3.3). It is guaranteed by the diffeomorphism generated by Hamiltonian vector field, $\boldsymbol{X}_H$, and it satisfies the detailed balance property of being ergodic.

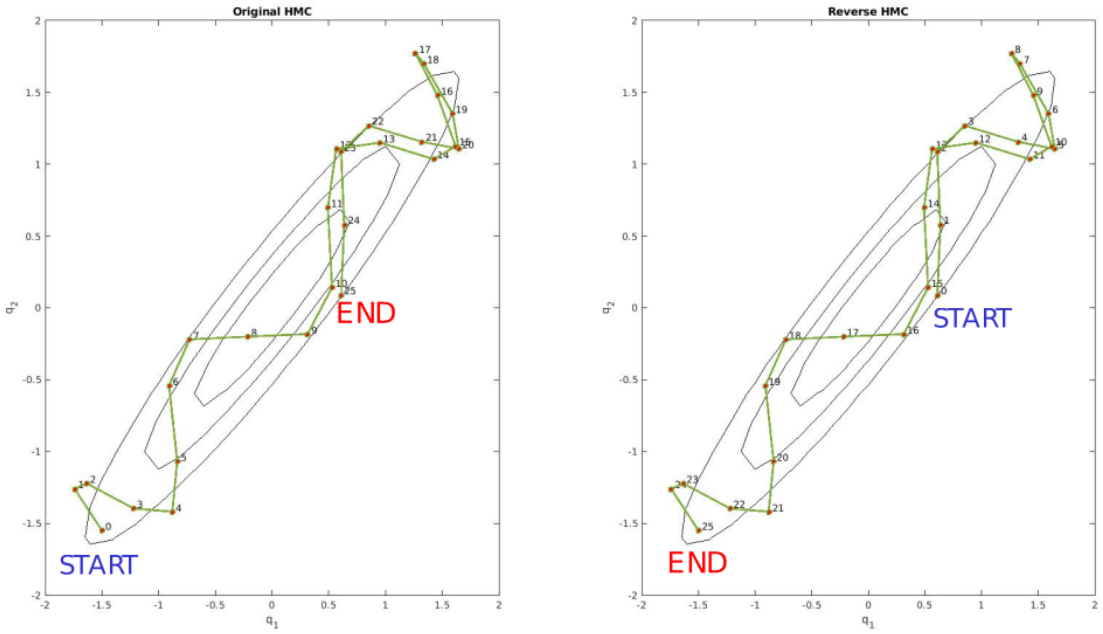

Figure 3.3. Illustration of the reversibility property in phase space.

Secondly, Hamiltonian dynamics preserves volume along the Hamiltonian trajectories by Liouville theorem 2.3, because the symplectic form $\boldsymbol{\omega}$ is preserved along the Hamiltonian vector field $\boldsymbol{X}_H$, i.e., $\mathcal{L}_{\boldsymbol{X}_H}\boldsymbol{\omega} = 0$. This property keeps the discrete desired distribution $\rho(\boldsymbol{q}, \boldsymbol{p})$ invariant in the phase space [16]. It gives us great convenience in the computational part, so that we can divide the phase space into equal volumes for numeric solution, hence we do not need to compute Jacobian factor for each sample.

Recall that the Hamiltonian dynamics guarantees preserving volume along the Hamiltonian vector field, i.e. sufficiently small volumes. In a normal distribution, even large volumes are preserved regardless of their size, as shown in Figures 3.4 and 3.6. On the other hand, the volume which is sufficiently small along the trajectories are preserved in any distribution. In non-symmetric distributions like non-Gaussian, volume is preserved even if its shape is not preserved as shown in Figure 3.5.

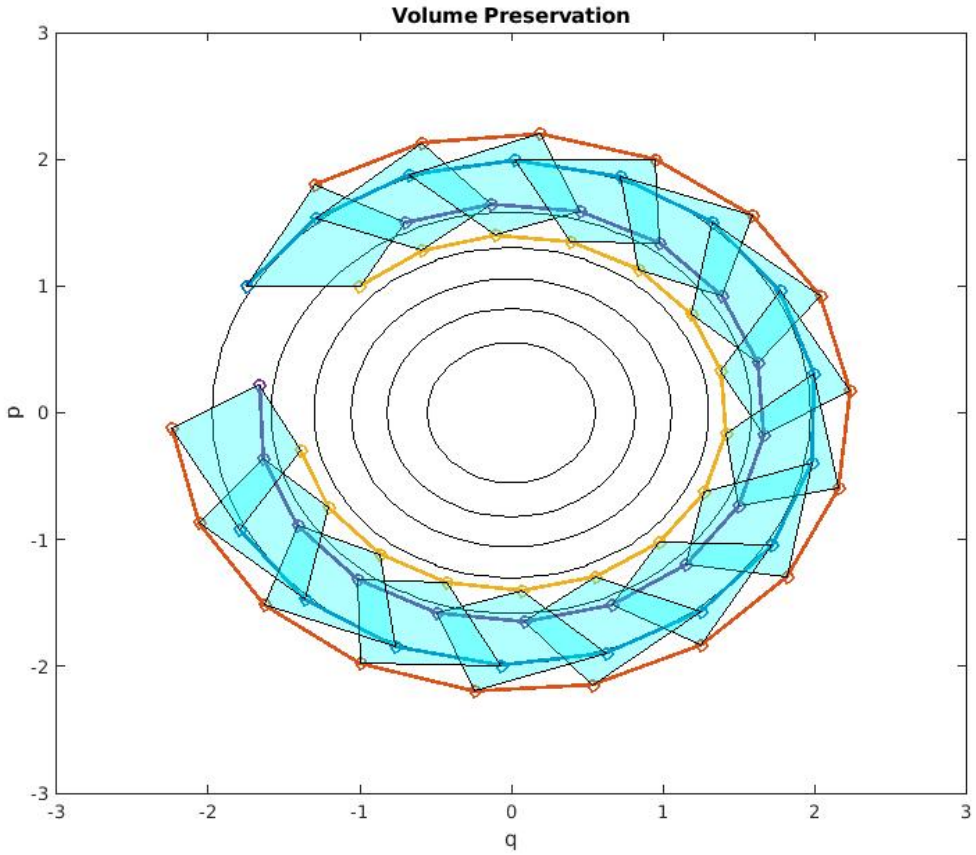

Figure 3.4. Illustration of the volume preservation property in phase space for a Gaussian distribution.

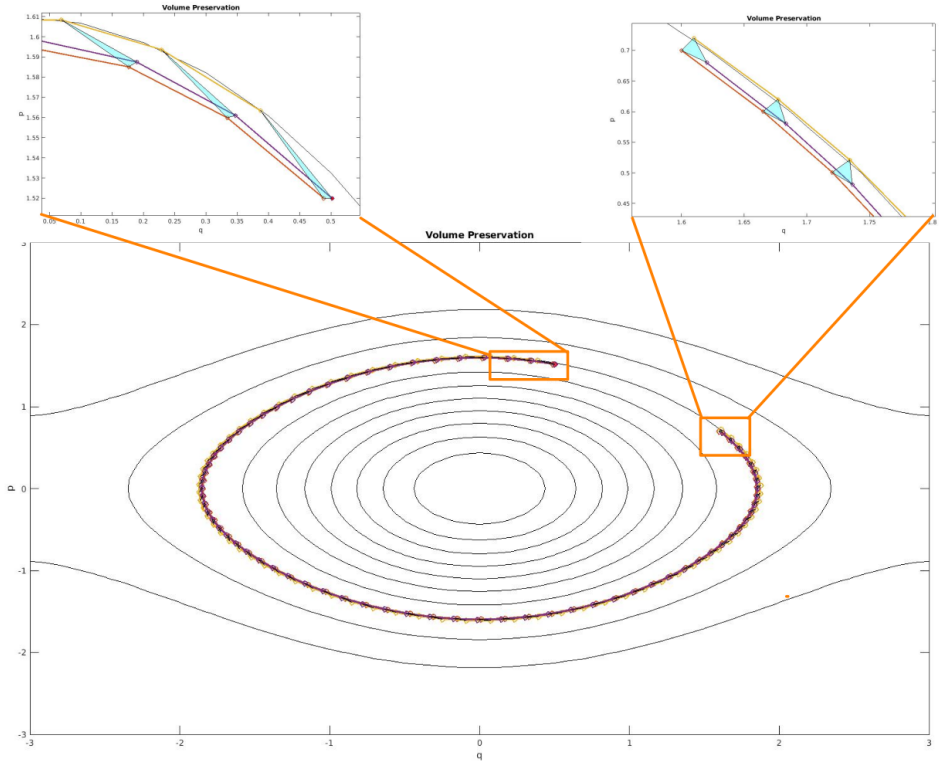

Figure 3.5. Illustration of the volume preservation property in phase space for a non-Gaussian distribution.

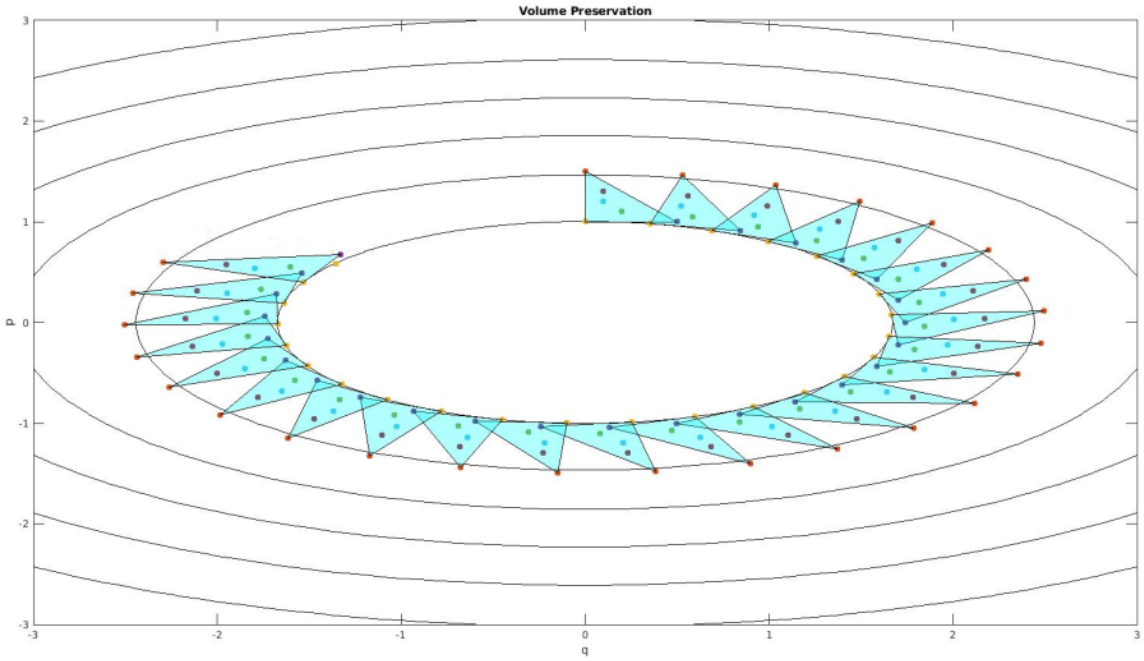

Figure 3.6. Even if the shape of the region changes, the area and the states inside of area are preserved.

Finally, the Hamiltonian function remains constant during the trajectory, since the derivative of the Hamiltonian with respect to time is zero. In application of HMC, it has some small numeric errors as in shown in Figure 3.7. This property is used during accept-reject test which is explained comprehensively in Section 4.2. Since Hamiltonian is invariant, acceptance probability is 1 without numerical error. We can get a new state which is distant from the current state with a high probability of acceptance thanks to the last property. In other words, HMC is able to establish the independence of the samples via tuning parameters properly.

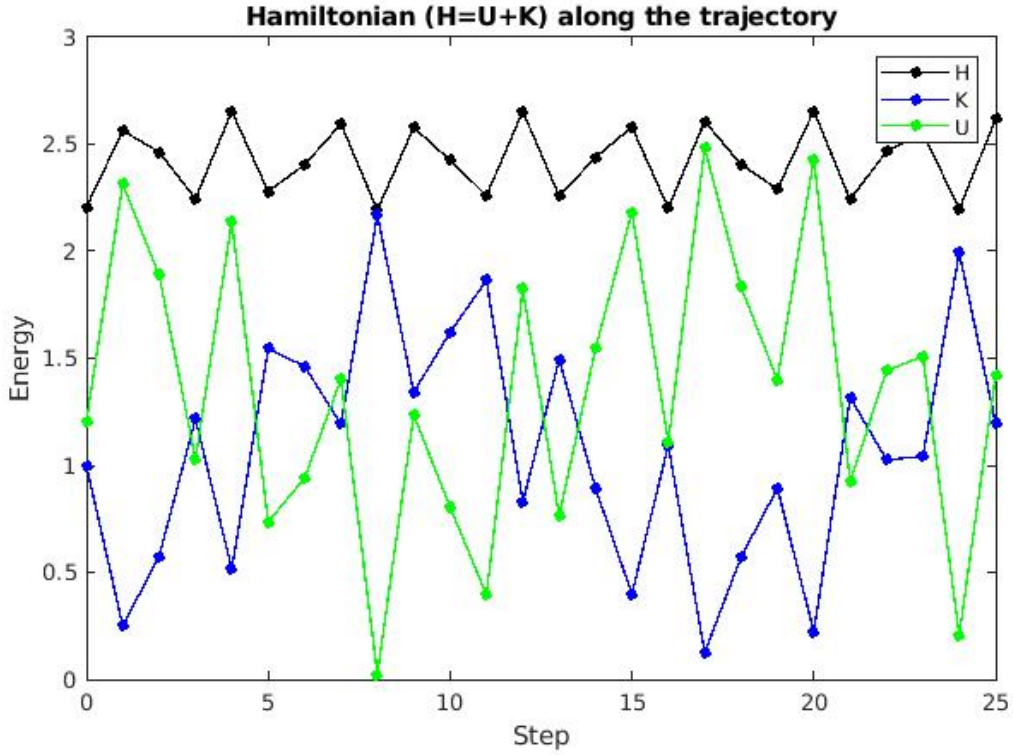

Figure 3.7. Illustration of the value of Hamiltonian function along the trajectory with small numeric error.

To summarize, HMC method is based on Hamiltonian dynamics. The properties of Hamiltonian dynamics take their power from phase space, accordingly symplectic geometry. This allows us to construct an ergodic Markov chain in state space at the end of the sampling.

4. HAMILTONIAN MONTE CARLO ALGORITHM

Basically, HMC is a sampling algorithm in which random samples are chosen according to a target probability distribution. The target distribution equals to the joint distribution, the product of posterior distribution and the distribution of momentum variable,

$$\rho(\boldsymbol{q}, \boldsymbol{p}) = \rho(\boldsymbol{q} \mid \boldsymbol{d})\rho(\boldsymbol{p}) = k\rho(\boldsymbol{d} \mid \boldsymbol{q})\rho(\boldsymbol{q})\rho(\boldsymbol{p}), \tag{4.1}$$

where k is the normalizer and Bayes' theorem (Theorem 3.1) is used in the second equality, namely $\rho(\boldsymbol{q} \mid \boldsymbol{d}) = k\rho(\boldsymbol{d} \mid \boldsymbol{q})\rho(\boldsymbol{q})$.

The joint distribution $\rho(\boldsymbol{q}, \boldsymbol{p})$ is actually a Gibbs canonical distribution, since it will be obtained from Hamiltonian function, i.e. $\rho(\boldsymbol{q}, \boldsymbol{p}) = \exp(-H(\boldsymbol{q}, \boldsymbol{p}))$. In other words, it is a function of energy of the system at a constant temperature.

In order to use Hamiltonian function for sampling, we need to combine Equation (2.24) and Equation (4.1) in a proper way. The potential energy of a sample is defined as the minus log of the posterior,

$$U(\boldsymbol{q}) = -\log(\rho(\boldsymbol{q} \mid \boldsymbol{d})), \tag{4.2}$$

where the posterior pdf is given by Bayes' theorem (Theorem 3.1),

$$\rho(\boldsymbol{q} \mid \boldsymbol{d}) \propto \rho(\boldsymbol{d} \mid \boldsymbol{q})\rho(\boldsymbol{q}), \tag{4.3}$$

where the likelihood, $\rho(\boldsymbol{d} \mid \boldsymbol{q})$, contains the forward relation between model and data spaces, and model prior, $\rho(\boldsymbol{q})$, contains all constraints on model space. As an example,

assuming Gaussian measurement error, one can obtain the likelihood function as

$$\rho(\boldsymbol{d} \mid \boldsymbol{q}) \propto exp\left(-\frac{1}{2}(\boldsymbol{d}_{obs} - g(\boldsymbol{q}))^T {C_D}^{-1}(\boldsymbol{d}_{obs} - g(\boldsymbol{q}))\right), \tag{4.4}$$

where $g(\boldsymbol{q})$ is the forward relation that gives the predicted data, and C_D is the data covariance matrix. Hence the potential energy corresponds to the L2-norm of the error between the observed data and the predicted data. Intuitively, the potential of a sample is high if it is far from the observed data, just like the potential energy of a ball increases as it rises from the ground.

On the other hand, we produce momentum artificially, because we need kinetic energy to move in the phase space. The momentum is defined as follows,

$$\boldsymbol{p} = M\dot{\boldsymbol{q}}, \tag{4.5}$$

where M is mass matrix in the Hamiltonian dynamics. Geometrically, M is a Riemann metric and momentum is a covector (one-form, dual of $\dot{\boldsymbol{q}}$) which is obtained by the multiplication of Riemann metric and velocity vector. If the momentum is considered as a one-form, it is a real valued function which eats a vector, by definition. Then, kinetic energy can be defined as

$$K(\dot{\boldsymbol{q}}) = \frac{1}{2}\boldsymbol{p}(\dot{\boldsymbol{q}}),$$

where $1/2$ is a normalization constant. On the other hand, we describe the momentum as a vector in calculation, so that kinetic energy is defined as

$$K(\boldsymbol{p}) = \frac{1}{2}\sum_{i,j=1}^{n} p_i M_{ij}^{-1} p_j. \tag{4.6}$$

Hence the marginal pdf of momentum can be defined as

$$\rho(\boldsymbol{p}) \propto exp(-K(\boldsymbol{p})) = \exp\left(-\frac{1}{2}\sum_{i,j=1}^{n} p_i M_{ij}^{-1} p_j\right), \tag{4.7}$$

which means that momentum $\boldsymbol{p}$ can be drawn from a Gaussian distribution.

Finally, the joint distribution in Equation (4.1) can rewritten with respect to the energy as follows

$$\rho(\boldsymbol{q},\boldsymbol{p}) = \exp(-U(\boldsymbol{q})\exp(-K(\boldsymbol{p})) = \exp(-U(\boldsymbol{q}) - K(\boldsymbol{p})) = \exp(-H(\boldsymbol{q},\boldsymbol{p})). \tag{4.8}$$

Observe that the canonical distribution is a Gibbs distribution in the phase space, since it is a function of energy of the Hamiltonian system with constant (ignored) temperature parameter. Choosing random samples according to the canonical distribution means moving in the phase space. Moreover, this movement is not exactly a random walk, because the movement is along a trajectory on which the Hamiltonian is conserved. That is so crucial for independence of samples and effective sampling with a high acceptance rate.

In this section, we are going to describe how to set up an Hamiltonian system on an inversion problem and how to write the corresponding algorithm.

4.1. Main Stages of Hamiltonian Monte Carlo

HMC algorithm requires three main stages; the first is to choose an initial state $(\boldsymbol{q}_0, \boldsymbol{p_0})$ whose components are, position (model) $\boldsymbol{q}_0$ and momentum $\boldsymbol{p}_0$. The position $\boldsymbol{q}_0$ might be any initial model, while the parameters of $\boldsymbol{p}_0$ are drawn from a Gaussian distribution $\mathcal{N}(0,1)$. The second main stage is to move in the phase space from the current state $(\boldsymbol{q}_0, \boldsymbol{p}_0)$ to a proposed state according to the Hamiltonian equations, i.e. along the Hamiltonian trajectory which is numerically solved. The last stage is to

make an accept-reject test at the end of the trajectory. The HMC algorithm works by repeating these three main stages iteratively.

Since the momentum is drawn from Gaussian distribution, the kinetic energy of the state is defined as in Equation (4.6),

$$K(\boldsymbol{p}) = \frac{1}{2}\sum_{i,j=1}^{n} p_i M_{ij}^{-1} p_j, \tag{4.9}$$

where M is mass matrix (Riemann metric, geometrically). Furthermore, the potential energy is determined by the minus log of posterior distribution (Equation (4.2)).

Although the aim is to sample from the posterior, we need to define a new distribution which is invariant in phase space, namely canonical distribution $\rho(\boldsymbol{q}, \boldsymbol{p})$, and then to achieve posterior by ignoring the momentum parameters or by integrating the canonical distribution over the momentum parameters.

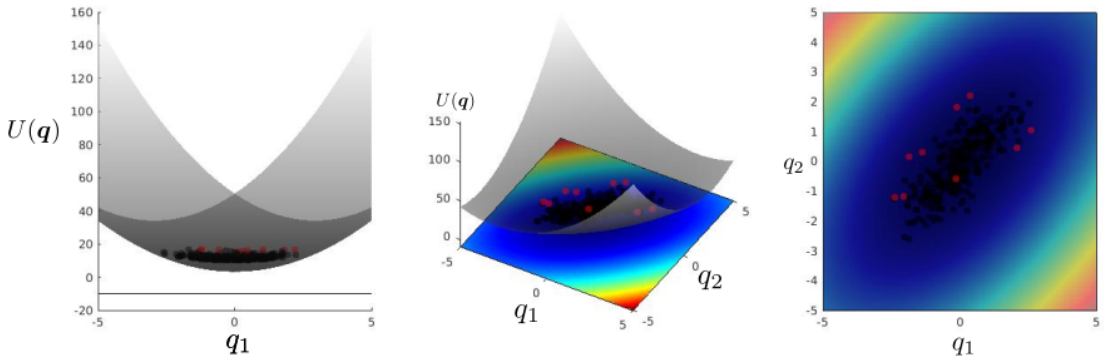

Figure 4.1. Illustration of sampling according to the posterior distribution (Equation (4.2)), red points are rejected and black points are accepted of 300 samples.

To summarize, the canonical distribution which is the stationary distribution of our Markov chain is defined as

$$\begin{aligned}\rho(\boldsymbol{q},\boldsymbol{p}) &= exp(-H(\boldsymbol{q},\boldsymbol{p})) \\ &= \exp(-U(\boldsymbol{q}) - K(\boldsymbol{p})) \\ &= \exp(-U(\boldsymbol{q}))\exp(-K(\boldsymbol{p})) \\ &= \rho(\boldsymbol{q} \mid \boldsymbol{d})\rho(\boldsymbol{p}),\end{aligned} \tag{4.10}$$

where $\rho(\boldsymbol{q} \mid \boldsymbol{d})$ is the posterior probability of $\boldsymbol{q}$ and $\rho(\boldsymbol{p})$ is the marginal probability of $\boldsymbol{p}$. This is the equation where Hamiltonian dynamics and Bayes' theorem meet for sampling.

4.2. Acceptance Probability

HMC algorithm collects samples according to the properties of Hamiltonian dynamics which is explained in Section 3.4. During application, these properties are not satisfied exactly because of some numeric errors, e.g. Hamiltonian is not exactly invariant as shown in Figure 3.7. Therefore, the proposed models during sampling should be tested to obtain a stationary distribution at the end of the sampling. By Theorem 3.3, detailed balance implies the existence of the stationary distribution. Considering the detailed balance property, the acceptance probability $A[(\boldsymbol{q}_1,\boldsymbol{p}_1)]$ of the proposed model at the end of the trajectory is determined as

$$A[(\boldsymbol{q}_1,\boldsymbol{p}_1)] = \min\left(1, \frac{\rho(\boldsymbol{q}_1,\boldsymbol{p}_1)}{\rho(\boldsymbol{q}_0,\boldsymbol{p}_0)}\right) = \min\left(1, \exp(-H(\boldsymbol{q}_1,\boldsymbol{p}_1) + H(\boldsymbol{q}_0,\boldsymbol{p}_0))\right) \tag{4.11}$$

The proposed state $(\boldsymbol{q}_1,\boldsymbol{p}_1)$ is accepted or rejected according to the acceptance probability. If it is accepted then the new Hamiltonian trajectory starts from $(\boldsymbol{q}_1,\boldsymbol{p}_1)$ (to go to new proposed state $(\boldsymbol{q}_2,\boldsymbol{p}_2)$). If the state $(\boldsymbol{q}_1,\boldsymbol{p}_1)$ is rejected, then the current step stays the same $(\boldsymbol{q}_0,\boldsymbol{p}_0)$. This process is repeated until sufficient number of samples are collected. Hence the accepted models form an ergodic Markov chain.

Observe from Equation (4.11) that if Hamiltonian equations can be solved analytically then the acceptance probability is always equal to unity, since Hamiltonian function is invariant. However, for an inversion process this is commonly not possible; problems are solved by numerically with some numerical errors, i.e. with changing acceptance probabilities.

How to obtain the Acceptance Probability, as a proof of Equation (4.11)

Numerically solving Hamiltonian equations requires to divide the phase space into some discrete equal volumes. In HMC, these volumes are determined as sufficiently small partitions, tangent space $\mathbb{X}_i$ at the state $(\boldsymbol{q}_i, \boldsymbol{p}_i)$. Therefore, the discrete probability can be written in terms of the density of the canonical distribution as $P(\mathbb{X}_i) = \mathbb{X}_i \rho(\boldsymbol{q}_i, \boldsymbol{p}_i)$. By Equation (3.7), the detailed balance between the current state $(\boldsymbol{q}_i, \boldsymbol{p}_i) \in \mathbb{X}_i$ and the proposed state $(\boldsymbol{q}_{i+1}, \boldsymbol{p}_{i+1}) \in \mathbb{X}_{i+1}$ on the same Hamiltonian trajectory is written as

$$
\begin{aligned}
T(\mathbb{X}_{i+1} \mid \mathbb{X}_i) P(\mathbb{X}_i) &= T(\mathbb{X}_i \mid \mathbb{X}_{i+1}) P(\mathbb{X}_{i+1}) \\
T(\mathbb{X}_{i+1} \mid \mathbb{X}_i)\ \mathbb{X}_i\ \rho(\boldsymbol{q}_i, \boldsymbol{p}_i) &= T(\mathbb{X}_i \mid \mathbb{X}_{i+1})\ \mathbb{X}_{i+1}\ \rho(\boldsymbol{q}_{i+1}, \boldsymbol{p}_{i+1})
\end{aligned}
$$

By the volume preservation property along the trajectory,

$$
\begin{aligned}
T(\mathbb{X}_{i+1} \mid \mathbb{X}_i)\ \rho(\boldsymbol{q}_i, \boldsymbol{p}_i) &= T(\mathbb{X}_i \mid \mathbb{X}_{i+1})\ \rho(\boldsymbol{q}_{i+1}, \boldsymbol{p}_{i+1}) \\
\frac{T(\mathbb{X}_{i+1} \mid \mathbb{X}_i)}{T(\mathbb{X}_i \mid \mathbb{X}_{i+1})} &= \frac{\rho(\boldsymbol{q}_{i+1}, \boldsymbol{p}_{i+1})}{\rho(\boldsymbol{q}_i, \boldsymbol{p}_i)}
\end{aligned}
$$

Hence, to satisfy the detailed balance property, the acceptance rate should be defined as

$$
\begin{aligned}
A[(\boldsymbol{q}_{i+1}, \boldsymbol{p}_{i+1})] &= \min\left(1, \frac{T(\mathbb{X}_{i+1} \mid \mathbb{X}_i)}{T(\mathbb{X}_i \mid \mathbb{X}_{i+1})}\right) \\
&= \min\left(1, \frac{\rho(\boldsymbol{q}_{i+1}, \boldsymbol{p}_{i+1})}{\rho(\boldsymbol{q}_i, \boldsymbol{p}_i)}\right) \\
&= \min\left(1, \exp(-H(\boldsymbol{q}_{i+1}, \boldsymbol{p}_{i+1}) + H(\boldsymbol{q}_i, \boldsymbol{p}_i)\right)
\end{aligned}
$$

Intuitively, if the canonical probability of proposed model is greater than the current model, then the proposed model will be accepted. In other words, if the Hamiltonian value (total energy) of the proposed model is less than the one of the current model, then the proposed model is definitely accepted. Otherwise, it will be rejected, depending on defined error margin.

4.3. Leapfrog Integrator

In order to solve the Hamiltonian equations numerically (to take steps along the Hamiltonian trajectories) it is necessary to maintain the detailed balance, which is satisfied by time reversibility and the volume preservation properties. The integrators which satisfy these properties are called *symplectic integrator*, and their characteristic feature is to preserve the symplectic form $d\boldsymbol{\omega} = d\boldsymbol{q} \wedge d\boldsymbol{p}$. One of the most common used symplectic integrators for HMC algorithms is the *leapfrog method*. It provides to walk from the current state to the proposed state along the trajectory, numerically. For an artificial time t and tuned parameter step size ε, the leapfrog method can be schematized as a first order Taylor expansion,

$$\begin{aligned}
\boldsymbol{p}_i\left(t+\frac{\varepsilon}{2}\right) &= \boldsymbol{p}_i + \frac{\varepsilon}{2}\frac{d\boldsymbol{p}_i}{dt}(t) \\
\boldsymbol{q}_i(t+\varepsilon) &= \boldsymbol{q}_i(t) + \varepsilon\frac{\boldsymbol{q}_i}{dt}(t) \\
\boldsymbol{p}_i\left(t+\varepsilon\right) &= \boldsymbol{p}_i\left(t+\frac{\varepsilon}{2}\right) + \frac{\varepsilon}{2}\frac{d\boldsymbol{p}_i}{dt}\left(t+\frac{\varepsilon}{2}\right).
\end{aligned} \tag{4.12}$$

By substituting the Hamiltonian equations (Equation (2.12)) into this scheme, the leapfrog is written as follows

$$\boldsymbol{p}_i\left(t+\frac{\varepsilon}{2}\right)=\boldsymbol{p}_i-\frac{\varepsilon}{2}\frac{\partial H}{\partial \boldsymbol{q}_i}(\boldsymbol{q}(t))$$

$$\boldsymbol{q}_i(t+\varepsilon)=\boldsymbol{q}_i(t)+\varepsilon\frac{\partial H}{\partial \boldsymbol{p}_i}\left(\boldsymbol{p}_i\left(t+\frac{\varepsilon}{2}\right)\right) \tag{4.13}$$

$$\boldsymbol{p}_i(t+\varepsilon)=\boldsymbol{p}_i\left(t+\frac{\varepsilon}{2}\right)-\frac{\varepsilon}{2}\frac{\partial H}{\partial \boldsymbol{q}_i}(\boldsymbol{q}_i(t+\varepsilon))$$

Finally, since the Hamiltonian function is defined as $H(\boldsymbol{q},\boldsymbol{p})=U(\boldsymbol{q})+K(\boldsymbol{p})$ where $K(\boldsymbol{p})=\dfrac{\boldsymbol{p}^T M^{-1}\boldsymbol{p}}{2}$ in our case, the form of the leapfrog is as follows,

$$\boldsymbol{p}_i\left(t+\frac{\varepsilon}{2}\right)=\boldsymbol{p}_i-\frac{\varepsilon}{2}\frac{\partial U}{\partial \boldsymbol{q}_i}(\boldsymbol{q}(t))$$

$$\boldsymbol{q}_i(t+\varepsilon)=\boldsymbol{q}_i(t)+\varepsilon M^{-1}\boldsymbol{p}_i\left(t+\frac{\varepsilon}{2}\right) \tag{4.14}$$

$$\boldsymbol{p}_i(t+\varepsilon)=\boldsymbol{p}_i\left(t+\frac{\varepsilon}{2}\right)-\frac{\varepsilon}{2}\frac{\partial U}{\partial \boldsymbol{q}_i}(\boldsymbol{q}_i(t+\varepsilon)).$$

In summary, the HMC algorithm for number of N iteration, written in MatLab can be shown as

```
%% HAMILTONIAN MONTE CARLO ALGORITHM FOR N SAMPLING

%% Initial state (q,p) in phase space
q0 = initial_position;
p0 = normrnd(0,1,[length(q0),1]); %drawn from the normal distribution

%% Start to walk in Phase Space
for k=1:N

    % Hamiltonian of the current state
    H(q0, p0) = U(q0) + K(p0);

    %% Tuning Parameters
    epsilon = step size;
    L = length of the trajectory

    % Artificial time t parameter w.r.t step size and length of the trajectory
    t = 0:epsilon:(L-1)*epsilon;

    %% Walking along the Hamiltonian Trajectory via Leapfrog method

    for i = 1:length(t)

        % Half step for momentum
        p_half = p0 - epsilon*grad_U(q0) / 2;

        % Full step for position until the end of the trajectory
        q_full = q + epsilon * inv(M) * p_half;

        % Second half step for momentum
        p_full = p_half - epsilon*grad_U(q_full) / 2;
    end

    %% The proposed state is at the end of the trajectory
    q_proposed = q_full;
    p_proposed = p_full;

    % Hamiltonian of the proposed state
    H(q_proposed, p_proposed) = U(q_proposed) + K(p_proposed);


    %% Accept- Reject Test

    if unifrnd(0,1) < exp(H0-H_proposed)
        q0 = q_proposed; %accept the proposed state
        p0 = normrnd(0,1,[length(q0),1]);
    else
        q0 = q0;  %reject the proposed state
        p0 = -p0;
    end
end
```

Figure 4.2. HMC Algorithm for number of N iterations

5. HAMILTONIAN MONTE CARLO METHOD FOR TRAVELTIME TOMOGRAPHY

Traveltime tomography is a process of inverting the velocity structure of a subsurface from the traveltime data of seismic waves. It is based on solving the eikonal equation, which describes the traveltime τ as a function for 2D velocity model $\mathbf{v}$,

$$|\nabla\tau|^2 = \frac{1}{\mathbf{v}^2(x,z)}. \tag{5.1}$$

In this chapter, it is created an artificial Hamiltonian system to estimate the velocity structure of a subsurface from the travel time data. Firstly, the corresponding forward relation is the Equation (5.1), where τ and $\mathbf{v}$ correspond the data parameter and the model, respectively. Regardless of inverting the velocity model from travel time problem, the forward relation can also be solved with Hamiltonian equations; where contours of Hamiltonian equations correspond to rays from source to receiver, as summarized in Table 2.1.

For a subsurface with 10 km depth and 30 km length, the subsurface can be gridded by 1 km such that

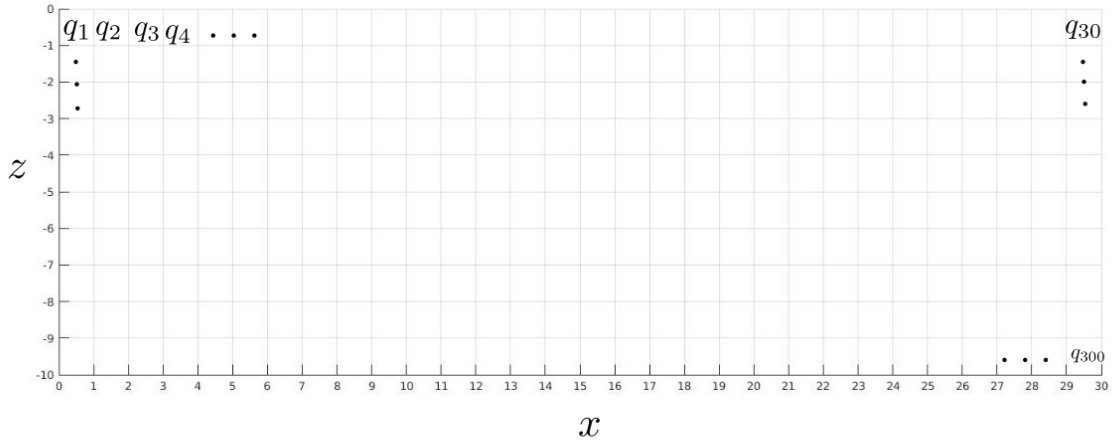

Figure 5.1. Subsurface profile with $10km \times 30km$, model parameters, symbolic rays between source and receiver

There exist 10×30 grid and each grid can have a different velocity in Figure 5.1, therefore the velocity model $\mathbf{v}$ in Equation (5.1) is 300 dimensional. Since the model corresponds to the position in Hamiltonian dynamics, position variable $\boldsymbol{q}$ is 300 dimensional, too. In order to invert the velocity structure by the Equation (5.1), sources and receivers are needed. The ray going from the source to the receiver will be the ray that goes in the fastest way in accordance with Snell and Fermat's law within the velocity structure. Hence the dimension of the travel time data $\boldsymbol{\tau}$ is equal to the multiplication of number of located earthquakes and the number of receivers in the subsurface, since we have a scalar traveltime value for each source and receiver pair, in Figure 5.2. Assume that there are 1 source and 30 receivers, hence our observed data is a 30-dimensional column vector.

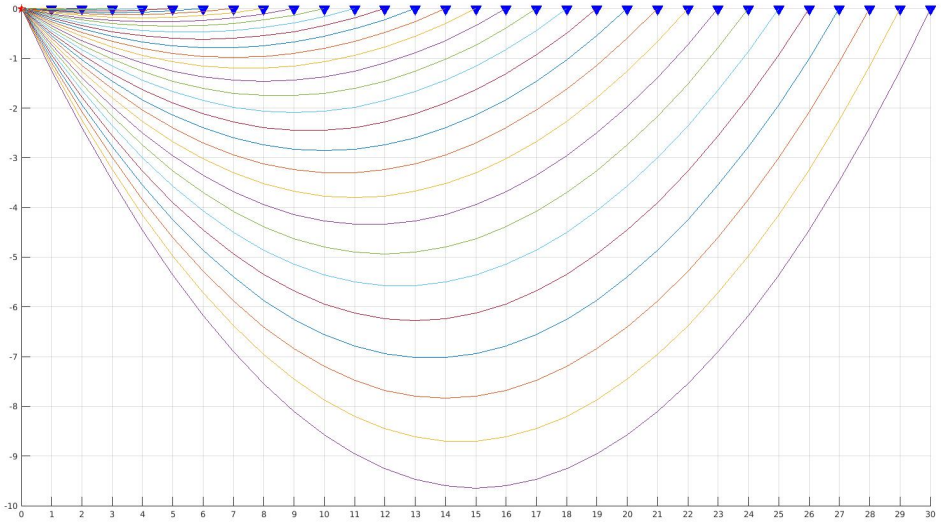

Figure 5.2. Symbolic rays between the source (red star) and receivers (blue triangles)

Traditionally, a traveltime tomography problem can be solved by gradient based methods; it is possible to establish a linear relationship between velocity and travel time increments and to solve it iteratively until achieving an acceptable misfit,

$$E(\mathbf{v}) = \frac{1}{2}(\boldsymbol{\tau}_{obs} - \boldsymbol{\tau}(\mathbf{v}))^T {C_D}^{-1}(\boldsymbol{\tau}_{obs} - \boldsymbol{\tau}(\mathbf{v})), \tag{5.2}$$

where C_D is data covariance matrix. The gradient based methods might be reasonable for weakly nonlinear cases, since a few iteration is sufficient for convergence. On the other hand, the velocity structure of a subsurface in the nature is quite complex for gradient based methods. In addition to the necessity of calculating the Jacobian matrix for each iteration, it requires to start with an initial model which is close to the solution. Otherwise, the iteration can get stuck in a local minimum region, which means bad solution. In addition to the possibility of not giving the best solution, it also brings high computation costs and long computational time. Therefore gradient-based methods are not preferred for a complex velocity structure. Hamiltonian Monte Carlo method, which is a probabilistic approach, provides remarkable solutions to most of

these problems, except calculating the gradient.

In HMC algorithm the misfit function corresponds to the potential energy of the current state;

$$U(\boldsymbol{q}) = \frac{1}{2}(\boldsymbol{d}_{obs} - \boldsymbol{\tau}(\boldsymbol{q}))^T {C_D}^{-1}(\boldsymbol{d}_{obs} - \boldsymbol{\tau}(\boldsymbol{q})), \tag{5.3}$$

where $\boldsymbol{\tau}(\boldsymbol{q})$ is the predicted data. As you can see in Figure 4.2, HMC requires to take gradient of potential, $U(\boldsymbol{q})$ during Leapfrog algorithm. By Chain rule, it is necessity to take gradient of the ray tracing function $\boldsymbol{\tau}(\boldsymbol{q})$. Therefore, it is crucial to have a well-defined ray-tracing code (forward relation) and an efficient gradient algorithm.

The power of HMC is based on the properties of phase space and symplectic form. The elements of phase space are the form of ordered pair of position and momentum $(\boldsymbol{q}, \boldsymbol{p})$. While position $\boldsymbol{q}$ corresponds to the model, momentum $\boldsymbol{p}$ is produced artificially to explore the state space. Therefore we need to define an artificial momentum $\boldsymbol{p}$ in order to move the problem into the phase space. In HMC methods, one can define kinetic energy in various ways [17]. In general, the kinetic energy is defined in its usual form as used in mechanics, Equation (4.6). Then, substituting this expression of kinetic energy to the Gibbs distribution, one obtains the Gaussian distribution.

Following an initial state, we can start a movement along the trajectory via leapfrog algorithm with suitable tuned stepsize ε and trajectory length L.

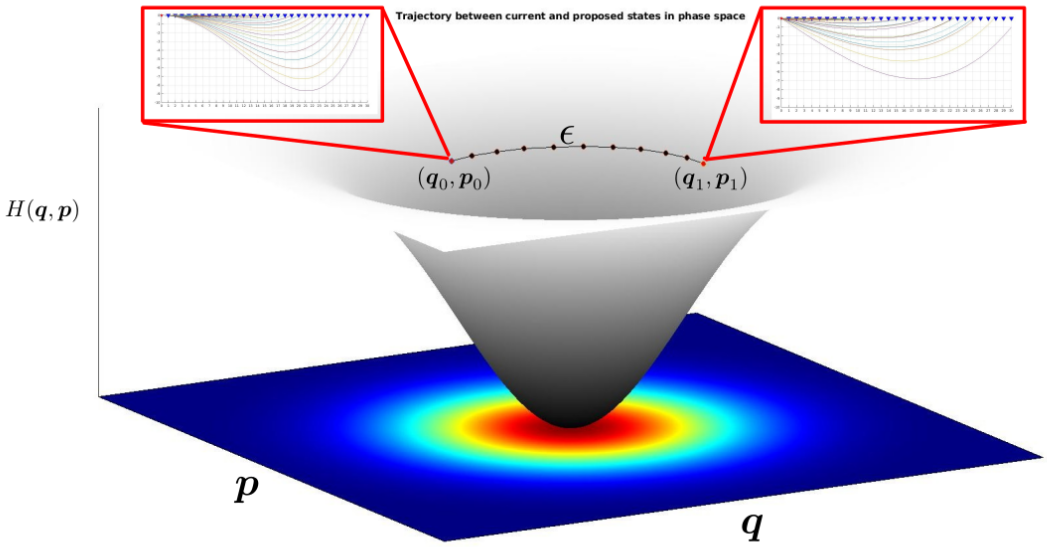

Figure 5.3. Symbolic first trajectory in phase space.

In Figure 5.3, the trajectory lies on the Hamiltonian contours. When the movement is projected onto the model space, s.t. when the momentum parameter $\boldsymbol{p}$ is ignored, we can see the movement in model space, As you can see in Figure 5.4 potential energy does not depend on momentum $\boldsymbol{p}$, where U is the misfit function which is form as Equation (5.3).

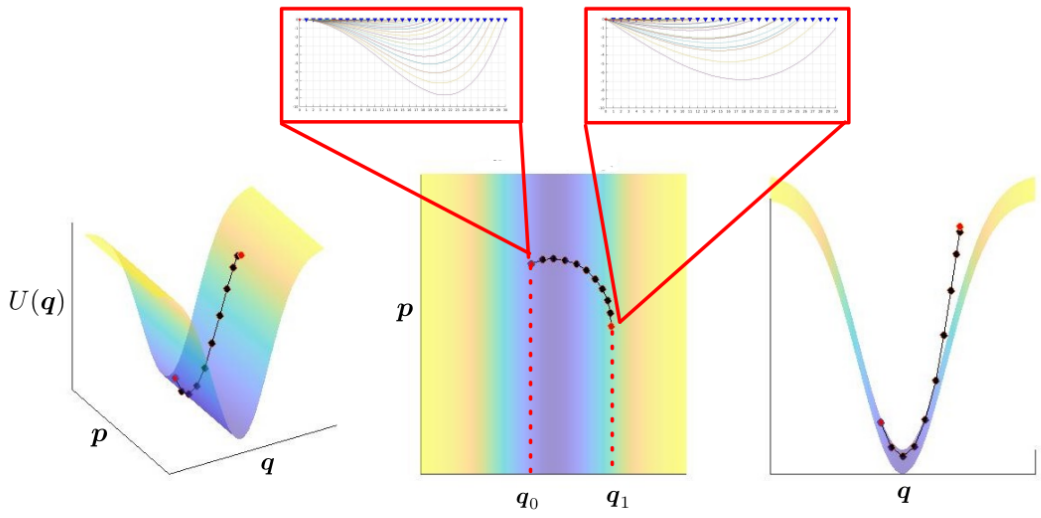

Figure 5.4. Symbolic first trajectory in model space in three different perspectives.

The state at the end of the trajectory is our proposed state and the accept-reject test is applied only on that point. As it is explained detailedly in Section 4.2, the accept-reject test is based on the stationary distribution which is obtained by Hamiltonian values of the state by Equation (4.11),

$$A[(\boldsymbol{q}_{i+1}, \boldsymbol{p}_{i+1})] = \min\left(1, \frac{\rho(\boldsymbol{q}_{i+1}, \boldsymbol{p}_{i+1})}{\rho(\boldsymbol{q}_i, \boldsymbol{p}_i)}\right) \tag{5.4}$$

$$= \min\left(1, \exp(-H(\boldsymbol{q}_{i+1}, \boldsymbol{p}_{i+1}) + H(\boldsymbol{q}_i, \boldsymbol{p}_i)\right), \; i \in \{0, \ldots, N\}.$$

If the proposed model is accepted, it is assumed as the current state and a new trajectory is started from that state. In Figure 5.5 and Figure 5.6, you can see the first three trajectories with accepted samples $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ in the model space. Furthermore, the accept-reject test will applied on the state $\boldsymbol{q}_3$.

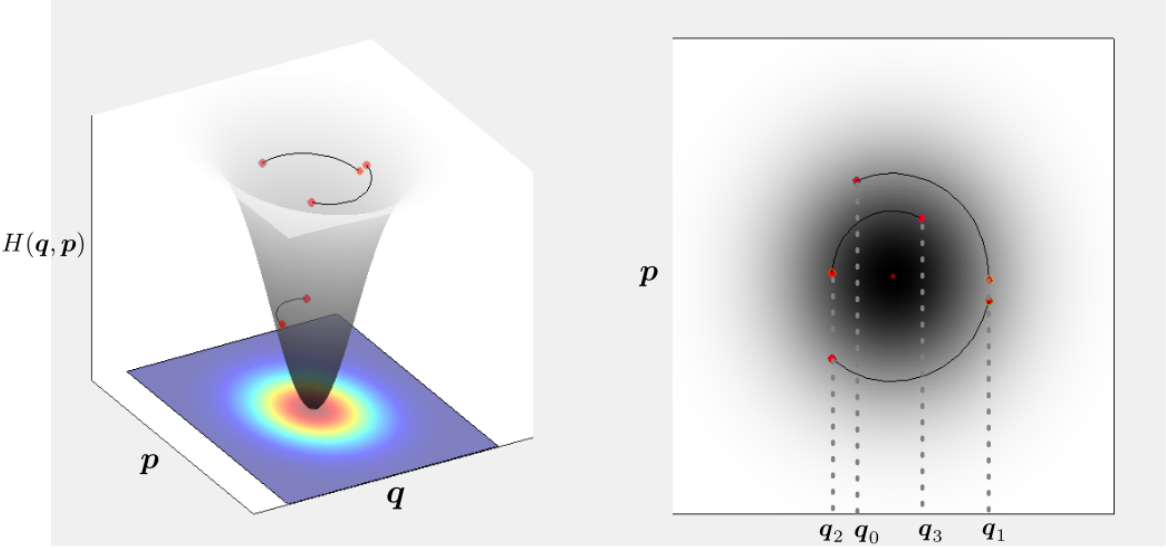

Figure 5.5. Symbolic 3 trajectories and their relations with canonical distributions.

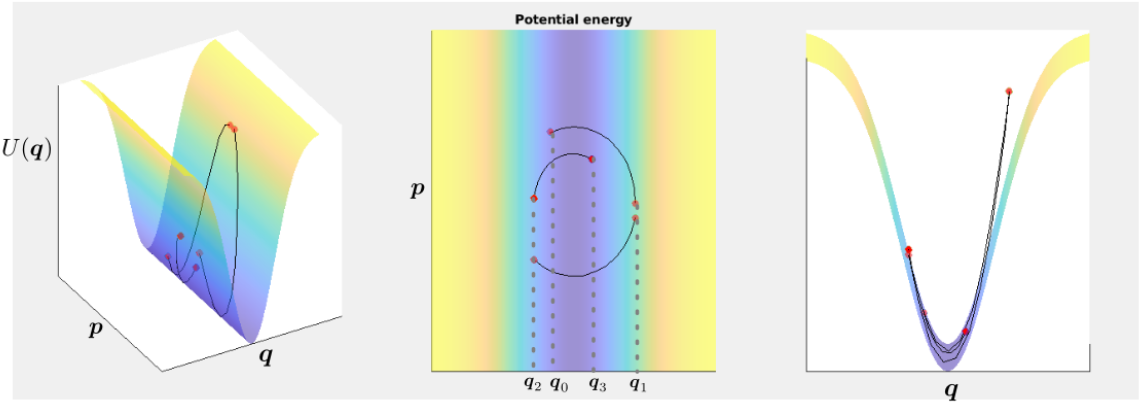

Figure 5.6. Symbolic 3 trajectories and their relation with likelihood distribution.

Surely, the best values for the tuned parameters ε and L depend on the problem and the model parameters. They must be determined neither too big nor too small. If ε is too big, then the trajectory does not lie along the Hamiltonian contours and accept-reject rate decrease, as shown in Figure 5.7.

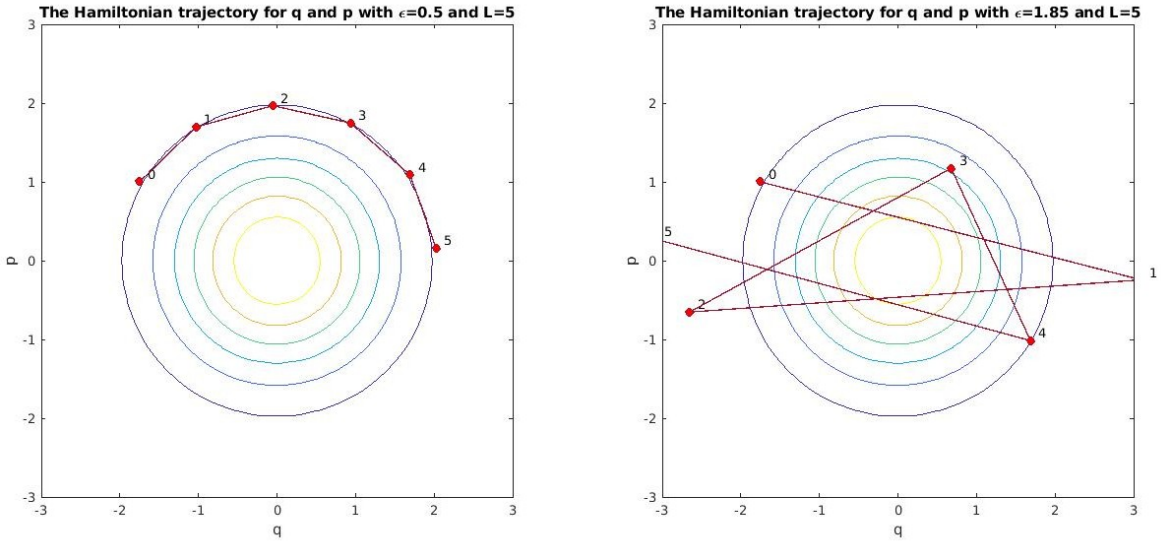

Figure 5.7. Illustration of tuning ε

Furthermore, if ε is too small, then it is required more sampling to explore model space. Moreover, if trajectory length L is too small or too big, it may not satisfy the independence of sampling. For example, in Figure 5.8, the distance between current model and the proposed model starts to decrease after ninth step, which means that $L = 9$ is the best for $\varepsilon = 0.3$.

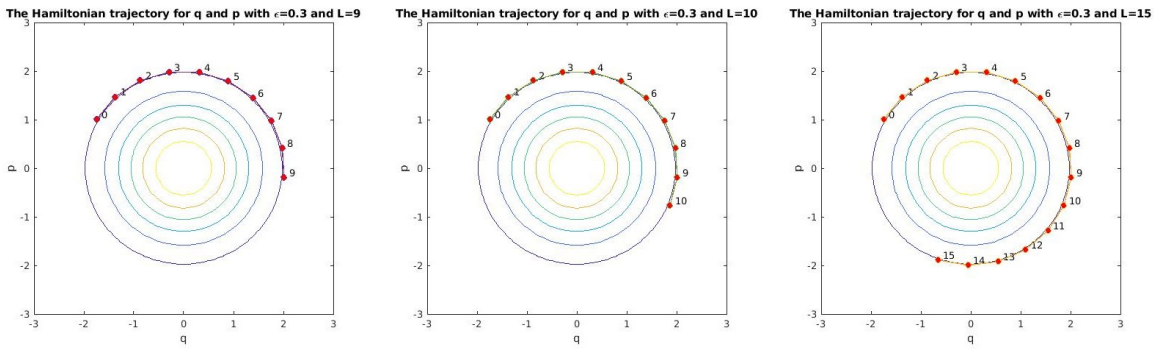

Figure 5.8. Illustration of tuning L

It is not possible to make visualize more then 3D space, therefore we cannot see the phase space of even 2D model space. However, it is convenient to show walking in the model space to see the continuity in model space where q^1 and q^2 are model parameters as shown in Figure 5.9.

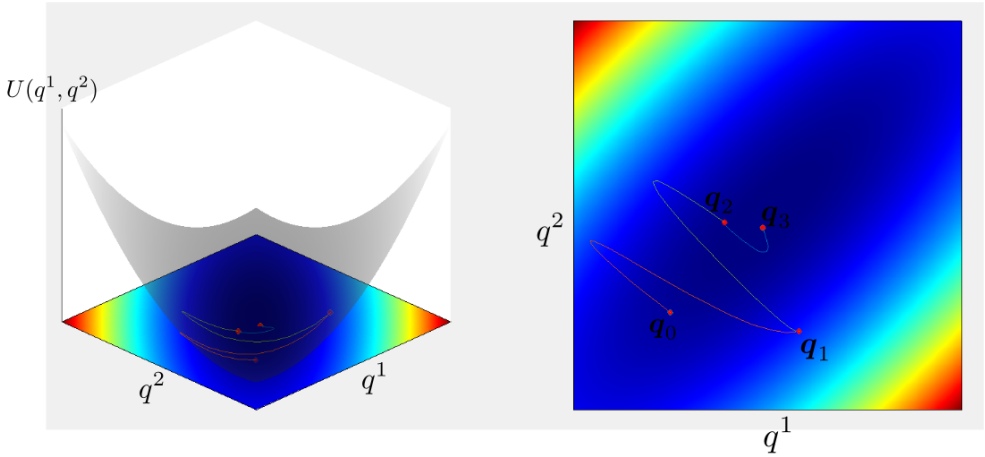

Figure 5.9. Walking in 2D model space, $\boldsymbol{q} = (q^1, q^2)$, and $\boldsymbol{q}_i$'s are samples.

Hence, we obtain an ergodic Markov Chain $\boldsymbol{q}_1, \boldsymbol{q}_2, ..., \boldsymbol{q}_N$ which satisfies detailed balance property. By Theorem 3.3, it has a stationary distribution. Thus, according to the Ergodic theorem (Theorem 3.2),

$$\mathbb{E}[\boldsymbol{q}] = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{q}_i \, \rho(\boldsymbol{q}_i \mid \boldsymbol{d}), \tag{5.5}$$

where N is the number of collected samples.

The expected value of the obtained Markov chain is an average solution of the tomography problem. It is also possible to check mode of the Markov chain if there is a repeated model. If the mean and mode are close to each other, our solution is most likely correct.

6. CONCLUSION

- In Chapter 2, we describe the relation between Hamiltonian Monte Carlo method and symplectic geometry. Hamiltonian dynamics, where symplectic geometry is originated from, is traditionally used to solve mechanics problems. Once the geometric relations of the phase space is revealed, Hamiltonian dynamics is started to be used in optimization problems, such as HMC.
- In Chapter 4, we explained the main stages of HMC algorithm. We emphasized that the acceptance probability is actually defined to provide detailed balance. Lastly, we introduced Leapfrog integrator which is used to walk along the trajectories. Finally, we wrote the HMC algorithm code for inversion.
- The use of HMC in travel-time tomography problems requires to work in several different mathematical spaces, which makes the problem looked complicated. In Chapter 5, we introduced all these spaces rigorously and describe the relation between them. More precisely, there are three spaces:
  - The domain of the forward function (eikonal equation) whose dimension is two for a 2D tomography problem, which means velocity structure varies with the horizontal (x) and vertical (z) components in the subsurface.
  - The position space (model space) whose dimension is the number of grid in subsurface (e.g. 300-dimensional in Figure 5.2). The coordinates of model parameters $\boldsymbol{q}$ corresponds to the seismic velocities in each grid.
  - The phase space whose dimension is twice the dimension of position space. The elements of the phase space are formed as $(\boldsymbol{q}, \boldsymbol{p})$.
- HMC method requires to tune parameters, namely step size ε and trajectory length L. They have a crucial importance for effectiveness of the algorithm.
  - For a fixed L, too big ε causes low acceptance rate due to the large error in Hamiltonian, while too small ε causes time-consuming.
  - For a fixed ε, too big L causes unnecessary calculations, while too small L does not guarantee the independence of samples and converges slowly.

- Hamiltonian Monte Carlo method have less computational cost due to the following reasons:
  - It does not walk randomly in the phase space; it walks along the Hamiltonian trajectories for sampling. Since Hamiltonian is invariant along these trajectories, the acceptance probability of proposed samples is high. It provides a faster convergence to solution with less sampling.
  - Requiring less samples implies taking fewer gradient.
  - There is no need to calculate the determinant of the Jacobian matrix for the volumes obtained for discretized pdf. We can ignore volume grids, because volume is preserved along the Hamiltonian trajectory by Liouville theorem.
- Hamiltonian Monte Carlo method has a few restriction in applicability to an inverse problem, mainly
  - It requires to calculate the predicted data for each proposed samples since the target distribution depends on the misfit between the observed and predicted data. Therefore, we should have a well-defined forward function which is preferred to be calculated fast.
  - It requires an efficient function for taking gradient fast, because it needs to take gradient at every step. Even for best optimized tuned parameters, the number of taking gradient is quite high, which increase the computational cost and time to compute.

REFERENCES

1. Backus, G. and F. Gilbert, "The Resolving Power of Gross Earth Data", *Geophysical Journal of the Royal Astronomical Society*, Vol. 16, 1968.

2. Backus, G. and F. Gilbert, "Uniqueness in The Inversion of Inaccurate Gross Earth Data", *Philosophical Transactions of the Royal Society*, Vol. 266, pp. 123–192, 1970.

3. Tarantola, A., *Inverse Problem Theory and Methods for Model Parameter Estimation*, Society for Industrial and Applied Mathematics, 2 edn., 2005.

4. Bayes, T. and R. Price, "An Essay Towards Solving a Problem in The Doctrine of Chance", *Philosophical Transactions of the Royal Society A*, Vol. 53, pp. 370–418, 1763.

5. Fichtner, A., A. Zunino and L. Gebraad, "Hamiltonian Monte Carlo Solution of Tomographic Inverse Problems", *Geophysical Journal International*, Vol. 216, pp. 1344–1363, 2019.

6. Keilis-Borok, V. and T. Yanovskaya, "Inverse Problems of Seismology (structural review)", *Geophysical Journal of the Royal Astronomical Society*, pp. 223–234, 1967.

7. Press, F., "Earth Models Obtained by Monte Carlo Inversion", *Journal of Geophysical Research*, pp. 5223–5234, 1968.

8. Sambridge, M., "Geophysical Inversion with the Neighbourhood Algorithm–I. Searching a Parameter Space", *Geophysical Journal International*, Vol. 138, pp. 479–494, 1999.

9. Sambridge, M., "Geophysical Inversion with the Neighbourhood Algorithm–II. Ap-

praising the Ensemble", *Geophysical Journal International*, Vol. 138, pp. 727–746, 1999.

10. Gallagher, K., M. S. Sambridge and G. G. Drijkoningen, "Genetic Algorithms: An Evolution of Monte Carlo Methods for Strongly Non-linear Geophysical Optimization Problems", *Geophysical Research Letters*, Vol. 18, pp. 2177—2180, 1991.

11. Betancourt, M., "A Conceptual Introduction to Hamiltonian Monte Carlo", *ArXiv*, 01 2017.

12. Neal, R., "MCMC Using Hamiltonian Dynamics", *Handbook of Markov Chain Monte Carlo*, pp. 113–162, 06 2012.

13. Wolpert, D. H. and W. G. Macready, "No Free Lunch Theorems for Optimization", *IEEE Transactions on Evolutionary Computation*, Vol. 1, pp. 67–82, 1997.

14. Betancourt, M. and L. Stein, "The Geometry of Hamiltonian Monte Carlo", *ArXiv*, 12 2011.

15. McInerney, A., *First Steps in Differential Geometry: Riemannian, Contact, Symplectic*, Springer, 1 edn., 2013.

16. Hand, L. N. and J. D. Finch, *Analytical Mechanics*, Cambridge University Press, Cambridge, 1998.

17. Livingstone, S., M. Faulkner and G. Roberts, "Kinetic Energy Choice in Hamiltonian/Hybrid Monte Carlo", *Biometrika*, Vol. 106, No. 2, pp. 303–319, 04 2019.