EFFICIENT MARKOV CHAIN MONTE CARLO METHODS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Youhan Fang

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2018

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Robert D. Skeel, Chair

Department of Computer Science

Dr. Ananth Grama

Department of Computer Science

Dr. Hisao Nakanishi

Department of Physics and Astronomy

Dr. Chunyi Peng

Department of Computer Science

Approved by:

Dr. Voicu Popescu by Dr. William J. Gorman

Head of the Graduate Program, Department of Computer Science

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABSTRACT

Fang, Youhan PhD, Purdue University, May 2018. Efficient Markov Chain Monte Carlo Methods. Major Professor: Robert D. Skeel.

Generating random samples from a prescribed distribution is one of the most important and challenging problems in machine learning, Bayesian statistics, and the simulation of materials. Markov Chain Monte Carlo (MCMC) methods are usually the required tool for this task, if the desired distribution is known only up to a multiplicative constant. Samples produced by an MCMC method are real values in N-dimensional space, called the configuration space. The distribution of such samples converges to the target distribution in the limit. However, existing MCMC methods still face many challenges that are not well resolved. Difficulties for sampling by using MCMC methods include, but not exclusively, dealing with high dimensional and multimodal problems, high computation cost due to extremely large datasets in Bayesian machine learning models, and lack of reliable indicators for detecting convergence and measuring the accuracy of sampling. This dissertation focuses on new theory and methodology for efficient MCMC methods that aim to overcome the aforementioned difficulties.

One contribution of this dissertation is generalizations of hybrid Monte Carlo (HMC). An HMC method combines a discretized dynamical system in an extended space, called the state space, and an acceptance test based on the Metropolis criterion. The discretized dynamical system used in HMC is volume preserving—meaning that in the state space, the absolute Jacobian of a map from one point on the trajectory to another is 1. Volume preservation is, however, not necessary for the general purpose of sampling. A general theory allowing the use of non-volume preserving dynamics for proposing MCMC moves is proposed. Examples including isokinetic

dynamics and variable mass Hamiltonian dynamics with an explicit integrator, are all designed with fewer restrictions based on the general theory. Experiments show improvement in efficiency for sampling high dimensional multimodal problems. A second contribution is stochastic gradient samplers with reduced bias. An in-depth analysis of the noise introduced by the stochastic gradient is provided. Two methods to reduce the bias in the distribution of samples are proposed. One is to correct the dynamics by using an estimated noise based on subsampled data, and the other is to introduce additional variables and corresponding dynamics to adaptively reduce the bias. Extensive experiments show that both methods outperform existing methods. A third contribution is quasi-reliable estimates of effective sample size. Proposed is a more reliable indicator—the longest integrated autocorrelation time over all functions in the state space—for detecting the convergence and measuring the accuracy of MCMC methods. The superiority of the new indicator is supported by experiments on both synthetic and real problems.

Minor contributions include a general framework of changing variables, and a numerical integrator for the Hamiltonian dynamics with fourth order accuracy. The idea of changing variables is to transform the potential energy function as a function of the original variable to a function of the new variable, such that undesired properties can be removed. Two examples are provided and preliminary experimental results are obtained for supporting this idea. The fourth order integrator is constructed by combining the idea of the simplified Takahashi-Imada method and a two-stage Hessian-based integrator. The proposed method, called two-stage simplified Takahashi-Imada method, shows outstanding performance over existing methods in high-dimensional sampling problems.

1 INTRODUCTION

One of the most important and challenging problems in machine learning, Bayesian statistics, and the simulation of materials is generating random samples from a prescribed distribution efficiently. In machine learning, for example, the size of data used for training models is of enormous size nowadays in the age of Big Data. The inference of probabilistic models, especially Bayesian models, requires extensive computation scaling up with the size of datasets. For this reason, maximum likelihood or maximum a posteriori estimations, which require only optimization procedures, are much more popular than full Bayesian treatments, which are based on sufficient number of samples of the posterior distribution. As is widely recognized, however, Bayesian treatments are in general more robust and more informative —providing the uncertainty on top of the estimates. Therefore, the usefulness and popularity of Bayesian models depend exclusively on algorithms that can generate random samples from the posterior distribution with comparatively low computation cost.

In many practical problems, the desired distribution is only known up to a multiplicative constant. In such cases, Markov Chain Monte Carlo (MCMC) methods, originally proposed by Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (MRRTT) in their landmark paper published in 1953 [1], are generally required. MCMC methods can generate unbiased samples by using a conditional acceptance criterion, usually called the Metropolis step. However, if based on random walk proposals, traditional MCMC methods generate highly correlated samples. Numerous steps are required to produce sufficiently independent samples; hence a great deal of computation power is wasted.

Hybrid Monte Carlo (HMC) methods, introduced in 1987 [2] and generalized in 1991 [3], combine dynamical systems with the Metropolis step and reduce correlation between successive samples. The dynamical systems are constructed to sample exactly

from the target distribution. They also use information from the gradient of the log density to reduce the random walk effect. The Metropolis step eliminates the bias due to discretization error introduced by the numerical integration of the dynamical system.

The dynamical system used in HMC is the Hamiltonian dynamics which has some special properties, such as "preserving volume"—meaning that the absolute Jacobian of a map from a point on the trajectory to another is 1. Such a special property, however, is unnecessary for the general purpose of sampling. In principle, any system of ordinary differential equations (ODEs) can be used for proposing new samples, as long as some certain conditions are satisfied. If the idea of HMC is extended to a more general framework, there can be many more possibilities.

There can be dynamical systems that are easier to transit from one mode to another in multimodal problems or that are more numerically stable. Thus, more general methods that use new dynamics and can be justified theoretically are needed.

An outstanding problem of dynamics based samplers is the potentially high computation cost in evaluating the gradient. In machine learning, for example, one evaluation of the gradient can be extremely expensive if the dataset is large. Methods based on stochastic gradients [4] have been very successful for reducing the cost of evaluating the gradient. A stochastic gradient is obtained by subsampling the data to approximate the true gradient. This idea is used in sampling methods based on stochastic differential equations (SDEs) such as Brownian dynamics [4, 5] and Langevin dynamics [6].

The cost of using stochastic gradients instead of full gradients is the introduction of distortion of the stationary distribution. The distortion is not harmless if not controlled. Therefore, it is of immense importance to find new methods that can reduce the distortion and, at the same time, keep the same computational efficiency as those naively using stochastic gradients.

Another important problem is to detect convergence and evaluate the accuracy of MCMC methods. Currently, this relies on computing the integrated autocorrelation

time (IAcT) of some particular functions and obtain the effective sample size (ESS) [7, 8]. The ESS is a quantity denoting the number of equivalent independent samples for a collection of dependent samples generated by the Markov chain in terms of some particular function whose expectation is to be estimated.

A small ESS implies the danger of incomplete sampling. Without sufficient samples, the estimated integrated autocorrelation time of some particular functions cannot be trusted. Of course, without additional information, it is impossible to find completely reliable method for detecting convergence. Nonetheless, by weakening the requirement—only considering apparent good coverage of state space—it is possible to define and estimate a more reliable quantity than the ordinary integrated autocorrelation time, to guarantee thorough sampling of those modes that have already been visited and to minimize the risk of missing an opening to yet another mode.

There are some other possibilities of improving the efficiency of sampling, under existing frameworks. For example [9], instead of designing new dynamical systems, consider changing the landscape of the potential energy function to avoid some difficulties, such as high energy barrier, in the original space.

Another example is to construct numerical integrators with higher order accuracy. Though simple, the classic leapfrog method has only order 2 accuracy. Higher order accuracy is desired without dramatically increasing the computation cost.

The aim of this dissertation is to propose methods and theory for solving aforementioned problems. Specifically, driven by desire to construct more efficient MCMC methods, dynamics-inspired methods more general than those based on Hamiltonian dynamics are sought. Limited practical success is achieved and interesting theoretical results are obtained. A different approach based on stochastic gradients is also pursued, again with some practical success. Doubts remain concerning the optimal choice of parameters and the actual efficiency of these methods, leading to a study of spectral gap and IAcT. A few additional attempts are made for improving the efficiency of samplers, and interesting results are obtained, suggesting new research directions.

To make the dissertation more readable, mathematical formulations are usually separated in the texts from motivation, description and discussion. In the remainder of Chap. 1, a brief introduction to the necessary mathematical background is presented in Sec. 1.1, and explicit statements of the challenges and contributions of the work are presented in Sec. 1.2.

1.1 Markov Chain Monte Carlo Methods

The goal of sampling is to generate identically distributed random samples from a target probability distribution. The probability density function can be written as

$$\rho(\theta) = \frac{1}{Z}\exp(-U(\theta)), \quad \theta = [\theta_1, \theta_2, \ldots, \theta_N],$$

where U is called the potential energy function, and Z is the normalizing constant usually unknown. In practical problems, $U(\theta)$ is usually high dimensional and has a non-trivial landscape that may contain many local minima.

The purpose of generating identically distributed samples from the target distribution is to estimate some quantity of interest $\mathbb{E}[u(\theta)]$, which is the expectation for specified function $u(\theta)$. Suppose the total number of samples is T, a quantity of interest is approximated by:

$$\mathbb{E}[u(\theta)] \approx \widehat{u} = \frac{1}{T}\sum_{i=0}^{T-1} u(\theta_i). \tag{1.1}$$

One example of the potential energy function is the Lennard-Jones potential, which is suitable for modeling interactions between the atoms of a noble gas. The pairwise potential function is

$$u(r_{ij}) = 4\varepsilon((\frac{\sigma}{r_{ij}})^{12} - (\frac{\sigma}{r_{ij}})^{6}),$$

where ε is the depth of the potential well, σ is the finite distance at which the inter-particle potential is zero, and r_{ij} is the distance between two particles, namely

$$r_{ij} = \sqrt{(\theta_{i1} - \theta_{j1})^2 + (\theta_{i2} - \theta_{j2})^2 + (\theta_{i3} - \theta_{j3})^2}.$$

The potential energy function for the entire system is

$$U(\theta) = \sum_{i,j,i<j} u(r_{ij}).$$

Note that energy units are chosen to eliminate the need for a temperature parameter to be consistent with statistical models.

In machine learning, the potential energy function is usually constructed from the likelihood function. In supervised learning, a data set consists of n examples. Each is a 2-tuple $(\mathbf{x}, y)$, where $\mathbf{x}$ represents the features of the observed example, and y is usually a scalar that represents the target value of the example (e.g., its label, output, activity, etc.). In unsupervised learning, each example has only the features $\mathbf{x}$. To make the notation general, in this dissertation, the data example is represented only by $\mathbf{x}$, while y appears only in specific supervised learning problems.

Let $\mathbf{X}$ denote the whole collection of data examples and $\rho(\mathbf{z})$ be the probability density function of random variables $\mathbf{z}$. The general likelihood function is

$$\rho(\mathbf{X}|\theta) = \prod_{i=1}^{n} \rho(\mathbf{x}_i|\theta).$$

For Bayesian models, there is a prior $\rho_0(\theta)$. $\rho(\theta|\mathbf{X})$ denotes the posterior distribution of θ given the data. The Bayesian rule gives

$$\rho(\theta|\mathbf{X}) = \frac{\rho(\mathbf{X}|\theta)\rho_0(\theta)}{\int \rho(\mathbf{X}|\theta)\rho_0(\theta)\mathrm{d}\theta}.$$

The denominator is the normalization constant equal to $\rho(\mathbf{X})$, which is also called the model evidence. The posterior probability density $\rho(\theta|\mathbf{X})$ can be also written as

$$\frac{1}{Z}\exp(-U(\theta)), \quad U(\theta) = -\log\rho(\mathbf{X}|\theta) - \log\rho_0(\theta),$$

with $Z = \rho(\mathbf{X})$.

One particular example in supervised learning is logistic regression [10], in which the logistic function maps a linear combination of features $\mathbf{x}$ to a probability:

$$\sigma(\theta; \mathbf{x}) = 1/(1 + \exp(-\theta^{\mathsf{T}}\mathbf{x})).$$

The potential energy function is

$$U(\theta) = \beta \sum_{i=1}^{n} (y_i \log(\sigma_i) + (1 - y_i) \log(1 - \sigma_i)) + \frac{1}{2}\alpha \|\theta\|^2 + \text{const.},$$

where α and β are weight parameters, and each y_i is either 1 or 0.

One may be interested in the expected value of the probability density of a new data example $\mathbf{x}$, given the training data $\mathbf{X}$. This can be written as

$$\rho(\mathbf{x}|\mathbf{X}) = \int \rho(\mathbf{x}|\theta)\rho(\theta|\mathbf{X})\mathrm{d}\theta.$$

Once one obtained the samples θ_i's from the posterior distribution $\rho(\theta|\mathbf{X})$, the quantity of interest can be estimated by

$$\rho(\mathbf{x}|\mathbf{X}) \approx \frac{1}{T} \sum_{i=0}^{T-1} \rho(\mathbf{x}|\theta_i).$$

1.1.1 The MRRTT Method

Consider a Markov chain

$$\theta_0 \to \theta_1 \to \theta_2 \to \cdots \to \theta_T.$$

The goal is to make the Markov chain converges to the target distribution $\rho(\theta)$, so that when T is large enough, θ_i's can be considered as identically distributed samples from the target distribution. Such samples are not in general independent. Nonetheless, since sampling is for estimating expectations, the degree of independence is only a consideration in terms of efficiency instead of correctness.

Convergence requires both stationarity and ergodicity. Stationary means $\theta_i \sim \rho(\theta) \Rightarrow \theta_{i+1} \sim \rho(\theta)$. Ergodicity means that any set with positive probability is visited by the chain almost surely [11] (Sec. II A).

The MRRTT method consists of two steps:

1. Given a sample θ, propose a move θ'.

2. Accept the proposed move θ' as the new sample with probability

$$\min\{1, \frac{\rho(\theta')}{\rho(\theta)}\},$$

otherwise, keep θ as the new sample.

This method guarantees stationarity.

Obviously, the quality of the proposal determines the efficiency. Longer moves and higher acceptance probabilities mean that the samples are less correlated. A traditional MRRTT method proposes samples based on a random walk and inevitably generates highly correlated samples.

1.1.2 The Hybrid Monte Carlo Method

HMC extends the configuration space of θ values to the phase space of values $\mathbf{z} = [\theta^\mathsf{T}, \mathbf{p}^\mathsf{T}]^\mathsf{T}$, where $\mathbf{p}$ is called the momenta. The joint density of θ and $\mathbf{p}$ can be written as $\rho(\theta, \mathbf{p}) \propto \exp(-H(\theta, \mathbf{p}))$ where $H(\theta, \mathbf{p}) = U(\theta) + \mathbf{p}^\top \mathbf{p}/2$ is the Hamiltonian. Each p_i in the momenta $\mathbf{p}$ follows the standard Gaussian distribution. The function $K(\mathbf{p}) = \mathbf{p}^\top \mathbf{p}/2$ is called the kinetic energy function, and the Hamiltonian is the total energy—sum of kinetic and potential energies. The force $\mathbf{f}(\theta) = -\nabla U(\theta)$ is the negative gradient of the potential energy. The Hamiltonian dynamics

$$\mathrm{d}\theta = \mathbf{p}\mathrm{d}t,$$
$$\mathrm{d}\mathbf{p} = \mathbf{f}(\theta)\mathrm{d}t,$$

maintains a constant total energy.

Let Ψ_τ^ν be the composition of ν numerical integrators $\Psi_{\Delta t}$ (e.g., the "leapfrog" method) of the Hamiltonian system, namely $\Psi_\tau^\nu = \Psi_{\Delta t} \circ \Psi_{\Delta t} \circ ... \circ \Psi_{\Delta t}$ and $\Delta t = \tau/\nu$. The HMC sampler can be described as follows:
Let θ be given.

1. Generate $\mathbf{p}$ from N independent standard Gaussian distribution

2. Compute $\mathbf{z}' = \Psi_\tau^\nu(\mathbf{z})$,

3. Accept θ' with probability

$$\min\{1, \frac{\rho(\mathbf{z}')}{\rho(\mathbf{z})}\}, \tag{1.2}$$

and otherwise keep θ.

One of the condition for HMC to satisfy stationarity [12] is that the propagator Ψ_τ^ν must preserve volume, meaning that the Jacobian $|\det(\partial\Psi_\tau^\nu/\partial\mathbf{z})| = 1$.

HMC explores the energy landscape much more efficiently than random walk MC for two reasons:

1. The Hamiltonian dynamics makes use of the information in the first order derivative of the potential energy function, such that the trajectory can find the region with high probability density much quicker than random walk.

2. The Hamiltonian dynamics conserves the total energy, or equivalently the joint density of θ and $\mathbf{p}$, along the trajectory, so that the acceptance probability for large moves can still be high even there is discretization error, if proper integrators are used.

1.1.3 Samplers Based on Stochastic Differential Equations

Another type of Markov chain samplers is based on the discretization of stochastic differential equations (SDEs). Usually, the Metropolis step is dropped in exchange for more efficient sampling. It is proved that there exists a modified stationary density for discretized SDEs, which is only slightly different (depending on the discretization step size Δt) from the the stationary distribution of continuous SDEs [13], and, in any case, bias due to discretization is usually dominated by statistical error. For this reason, SDE based samplers are widely used in molecular dynamics and machine learning. Moreover, in machine learning, such samplers are even a must, because the Metropolis step is incompatible with the use of stochastic gradients, which is discussed in detail later.

The Langevin dynamics is described by the following system of SDEs:

$$\begin{aligned} \mathrm{d}\theta &= \mathbf{p}\mathrm{d}t, \\ \mathrm{d}\mathbf{p} &= \mathbf{f}(\theta)\mathrm{d}t - A\mathbf{p}\mathrm{d}t + \sqrt{2A}\mathrm{d}\mathbf{w}, \end{aligned} \tag{1.3}$$

where A is a scalar and $\mathbf{w}$ is N independent Wiener processes. A Wiener process w satisfies:

1. $w(0) = 0$ with probability 1.

2. $w(t + \Delta t) - w(t) \sim \mathcal{N}(0, \Delta t)$ and is independent of $w(s)$ for $s \leq t$.

The stochastic term $\mathrm{d}w$ sometimes is informally written as $\mathcal{N}(0, \mathrm{d}t)$ [6].

By rescaling time $t \leftarrow At$ and letting $A \to \infty$, i.e., neglecting inertia effects on long time scales, one obtains the Brownian dynamics

$$\mathrm{d}\theta = \mathbf{f}(\theta)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{w}. \tag{1.4}$$

1.1.4 Stochastic Gradients

In machine learning, the gradient of the negative log likelihood function is written as

$$-\nabla \log \rho(\mathbf{X}|\theta) = -\sum_{i=1}^{n} \nabla \log \rho(\mathbf{x}_i|\theta).$$

When the size of the dataset is large, the computation cost of evaluating the gradient can be very high. In such cases, a random subset of data $\mathbf{x}_i$'s is sampled from the full data set, and an approximation of the gradient is obtained by using the subset of data.

Specifically, as stated in Sec. 1.1, the potential energy function of a Bayesian model is the negative logarithm of the probability density of the posterior distribution:

$$U(\theta) = -\sum_{i=1}^{n} \log \rho(\mathbf{x}_i|\theta) - \log \rho(\theta).$$

Let $U_i = -\log \rho(\mathbf{x}_i|\theta)$, and rewrite U to be

$$U(\theta) = \sum_{i=1}^{n} U_i(\theta) + U_0(\theta),$$

where the term U_0 representing the prior is henceforth omitted when n is large. Let the size of the random subset be m. The stochastic gradient, $\nabla\tilde{U}(\theta)$, is written as

$$\nabla\tilde{U}(\theta) = \sum_{i=1}^{n}(1 + r_i)\nabla U_i(\theta), \tag{1.5}$$

where r_i are correlated random variables, each one assuming the value -1 or $n/m-1$.

1.1.5 Integrated Autocorrelation Time and Effective Sample Size

As given in Eq. (1.1), the quantity of interest $\mathbb{E}[u(\theta)]$ is approximated by the average of the function u evaluated at the samples.

The variance of the estimated quantity is given by

$$\mathrm{Var}[\widehat{u}] = \frac{1}{T}\mathrm{Var}[u(\theta)]\left(1 + 2\sum_{i=1}^{T-1}\left(1 - \frac{i}{T}\right)\frac{C(i)}{C(0)}\right),$$

where the autocovariances

$$C(i) = \mathbb{E}[(u(\theta_0) - \mu)(u(\theta_i) - \mu)],$$

with $\mu = \mathbb{E}[u(\theta)]$. As $T \to \infty$,

$$\mathrm{Var}[\widehat{u}] = \frac{1}{T}\mathrm{Var}[u(\theta)]\tau + \mathcal{O}(\frac{1}{T^2}),$$

where τ is the *integrated autocorrelation time*.

$$\tau = 1 + 2\sum_{i=1}^{+\infty}\frac{C(i)}{C(0)}. \tag{1.6}$$

Note that if the samples are independent, the variance of the expected mean should be $\mathrm{Var}[u(\theta)]/T$. Therefore, the integrated autocorrelation time τ indicates the degree of dependence among the samples. The effective sample size is defined as

$$ESS = \frac{T}{\tau}. \tag{1.7}$$

In practice, when the ESS is large enough, say > 1000, the sampling is considered complete.

1.2 Challenges and Contributions

In this section, challenges of MCMC methods and contributions of this dissertation are presented.

1.2.1 Generalizations of Hybrid Monte Carlo

One of the main challenges for ordinary HMC is the metastability in multimodal problems. The trajectory can be trapped in one of the local minima of the potential energy function such that mixing is almost impossible in reasonable time. In addition, the Hamiltonian dynamics is not necessarily the best choice in terms of integrator step size, which is determined by numerical stability and accuracy considerations.

The idea is to broaden the possibilities for constructing proposals by extending HMC to a more general framework. In this dissertation, some theoretical results, which weaken the sufficient conditions for stationarity, are presented. Specifically, the condition of preserving volume— the Jacobian of the map defined based on the discretized dyanmics being 1— is removed, and reversibility is required only in the form of a bijection rather than an involution. Using the new framework, better dynamics can be designed with less restrictions.

Four examples that justify the theory are shown. All of the examples have improvement on ordinary HMC, by either crossing the energy barrier more frequently, or being more numerically stable or accurate—meaning more independent samples for the same computation cost.

1.2.2 Stochastic Gradient Samplers with Reduced Bias

In previous attempts [4–6] that utilize stochastic gradients for sampling, stochastic gradients merely replace the full gradients of SDEs, such as Langevin dynamics and Brownian dynamics. Although reducing the computation cost significantly, this idea makes the bias of the stationary distribution too large to be tolerated.

An in-depth analysis on the distortion caused by stochastic gradients is provided, and two methods to remove, at least partially, the damage to the stationary distribution caused by stochastic gradients are proposed. One method is based on the estimation of the noise and corrects the dynamics directly with these estimates, while the other is by the introduction of a new variable and corresponding dynamics that perform as a thermostat to remove the distortion automatically. Extensive numerical experiments are performed to show the superiority of the new methods.

1.2.3 Quasi-reliable Estimates of Effective Sample Size

In some difficult sampling problems, such as sampling from multimodal distributions, the integrated autocorrelation time of a particular function may not be trustworthy due to incomplete sampling. In addition, functions of interest that are popular in some applications may not be representative for the purpose of convergence detection —the best function is the eigenfunction corresponding to the eigenvalue of the propagator (a transfer operator of probabilities) closest to 1 and is in general unknown.

Introduced is the longest autocorrelation time $\tau_{\max}$ over all possible functions defined on state space. It is a more reliable indicator of convergence and a better measurement of the efficiency of Markov chain samplers, than the τ of just any particular function.

An algorithm is also proposed for estimating $\tau_{\max}$. The algorithm assumes only that a method for estimating τ for an arbitrary function is available. And its computation cost is much less than the cost of the sampling procedure. Numerical evidence is also presented for the utility of this algorithm.

1.2.4 Further Work on Sampling Methods

The difficulty of efficient sampling arises from certain undesired properties of the potential energy function, such as the existence of high energy barrier or extremely

elongated low energy basins. Inspired by the idea of spatial warping [9], a simple general formula for change of variables is presented. By changing variables, dynamical systems of the original variable are transformed into those of new variables, such that undesired properties of the potential energy function are removed. The new method is simpler and requires less problem specific knowledge than the spatial warping method. Two concrete examples are presented to illustrate the utility of this idea.

As a classic numerical integrator, the leapfrog method is the most widely used integrator for the Hamiltonian system. The order of accuracy of the leapfrog method is 2, meaning that the error is bounded by $\mathcal{O}(\Delta t^2)$. Proposed is a numerical integrator with accuracy of order 4, by combining the idea of the simplified Takahashi-Imada method [14] and a certain two-stage scheme [15]. Numerical experiments show the outstanding performance of the proposed method.

2 GENERALIZATIONS OF HYBRID MONTE CARLO

The traditional random-walk MCMC methods do not explore the energy landscape efficiently. Large moves usually result in low acceptance probability, and the trajectory of the Markov chain goes back and forth so frequently that a lot of computation is wasted. A hybrid Monte Carlo (HMC) method extends configuration space to phase space by introducing auxiliary variables, the momenta, and combines the Hamiltonian dynamics step for proposing a new sample and the Metropolis step compensating for the discretization error caused by numerical integrators.

An HMC method is able to produce samples from the target distribution [12], since it satisfies the sufficient conditions: (i) It is stationary, because the discretized (leapfrog method) Hamiltonian dynamics is reversible and volume preserving. (ii) It is ergodic, because the momenta are randomized in each MC step.

Also, an HMC method is efficient, for two reasons: (i) It has high acceptance probability in the Metropolis step. When the force is equal to the negative gradient of the potential energy, which is defined to be the negative logarithm of the desired probability density, the Hamiltonian dynamics keeps the value of the density function constant. This means that if the dynamics can be solved exactly, there is no change on the value of the density function, hence the acceptance probability of the Metropolis step is always 1. Of course, numerical integrators introduce discretization error, but for sufficiently small step size, the error is small, and the acceptance probability of the Metropolis step is close to 1. (ii) The auxiliary momenta variables contribute to the ballistic movement [16], which offers speedups for finding high probability regions, by following the gradient of the potential energy function.

Naturally, one may ask whether it is possible to use other dynamical systems, combining with the Metropolis step, to do sampling more efficiently? The answer is yes, as long as there is a theoretically justified framework (Sec. 2.2) for finding such

dynamics. There are two things to keep in mind when constructing such framework: (i) the dynamical system combined with the Metropolis step should be able to produce the desired density, (ii) the acceptance probability of the Metropolis step should be close to 1.

It is found that if the target density and the dynamics satisfy the continuity equation (Eq. (2.5)), which states that the negative divergence of the probability current equals the rate of change of the probability density, and the Metropolis criterion has a more generalized form—involving the Jacobian of the discretized dynamics, the aforementioned two goals can both be achieved. Given the target distribution, the framework is used to design new dynamics with good properties, such as crossing the energy barrier easier, which is made possible by removing the restriction of phase-space volume preservation. And at the same time, it is expected that the acceptance probability is as high as that of the traditional HMC. Four examples are presented to show the utility of the framework.

The remainder of this chapter is organized as follow: Sec. 2.1 reviews the basic idea of HMC and related topics; Sec. 2.2 presents the general framework; Sec. 2.3 shows four examples, two of which are further supported by numerical experiments; Sec. 2.4 concludes this chapter with a discussion.

2.1 Background

In HMC, the joint density of θ and $\mathbf{p}$ is $\rho(\theta, \mathbf{p}) \propto \exp(-H(\theta, \mathbf{p}))$ where $H(\theta, \mathbf{p}) = U(\theta) + \mathbf{p}^\top \mathbf{p}/2$. The Hamiltonian dynamics

$$\begin{aligned} \mathrm{d}\theta &= \mathbf{p}\mathrm{d}t, \\ \mathrm{d}\mathbf{p} &= \mathbf{f}(\theta)\mathrm{d}t, \end{aligned}$$

has this joint density as its invariant density. Note that the joint density has the marginal distribution equal to the target distribution, i.e.,

$$\int \rho(\theta, \mathbf{p})\mathrm{d}\mathbf{p} = \rho(\theta).$$

Use $\mathbf{z}$ to represent the extended space variable $\mathbf{z} = [\theta^\mathsf{T}, \mathbf{p}^\mathsf{T}]^\mathsf{T}$. A general dynamical system can be written as

$$\mathrm{d}\mathbf{z} = \mathbf{v}(\mathbf{z})\mathrm{d}t, \tag{2.1}$$

where $\mathbf{v}$ is the vector field that describes the motion of particles (characterized by positions θ and momenta $\mathbf{p}$) in the system.

The flow Φ_t is defined in terms of a vector field $\mathbf{v}(\mathbf{z})$ by a system of ODEs

$$\frac{\mathrm{d}\Phi_t}{\mathrm{d}t} = \mathbf{v} \circ \Phi_t, \quad \Phi_0 = \mathrm{id}.$$

Let $\mathbf{z}_t = \Phi_t(\mathbf{z}_0)$ be the flow map of $\mathbf{z}$ starting from $\mathbf{z}_0$. A flow map being *reversible* under the transformation R means

$$\Phi_t^{-1}(\mathbf{z}) = R(\Phi_t(R(\mathbf{z}))). \tag{2.2}$$

An example of R is $R : [\theta^\mathsf{T}, \mathbf{p}^\mathsf{T}]^\mathsf{T} \mapsto [\theta^\mathsf{T}, -\mathbf{p}^\mathsf{T}]^\mathsf{T}$. It is easy to see that the flow map of the Hamiltonian dynamics is reversible under this R. The Hamiltonian dynamics also preserves volume, since the vector field $\mathbf{v}$ is divergence free [17], i.e., $\nabla \cdot \mathbf{v} = 0$.

Note that the reversibility here is for the flow rather than the MCMC sampler. Reversibility for an MCMC sampler means that it satisfies *detailed balance*:

$$\rho_\mathrm{t}(\mathbf{z}'|\mathbf{z})\, \rho(\mathbf{z}) = \rho_\mathrm{t}(\mathbf{z}|\mathbf{z}')\, \rho(\mathbf{z}'), \tag{2.3}$$

where ρ_t denotes the transition density. Detailed balance is sufficient for stationarity.

The density function for the dynamical system is defined with respect to t and $\mathbf{z}$ as $\rho(t, \mathbf{z})$. The continuity equation is

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0. \tag{2.4}$$

By letting $\partial\rho/\partial t = 0$, it is easy to see that a probability density ρ is the invariant density of the dynamics described by $\mathbf{v}$, if the dynamics and the probability density satisfy the *stationary* continuity equation

$$\nabla \cdot (\rho \mathbf{v}) = 0. \tag{2.5}$$

Since non-stationary density is not in consideration in any cases in this dissertation, the stationary continuity equation is simply referred as the continuity equation, and ρ is referred as the stationary density which does not depend on t. Let $\rho \propto \exp(-H)$. The continuity equation can also be written as

$$\nabla H \cdot \mathbf{v} = \nabla \cdot \mathbf{v}. \tag{2.6}$$

The joint density ρ and the vector field $\mathbf{v}$ of the Hamiltonian dynamics satisfy the continuity equation. It is easy to see that the Hamiltonian H is a conserved quantity, namely, for two values $\mathbf{z}$ and $\mathbf{z}' = \Phi_t(\mathbf{z})$,

$$H(\mathbf{z}) = H(\mathbf{z}').$$

Recall that the Metropolis criterion of HMC given $\mathbf{z}$ and $\mathbf{z}'$ is

$$\min\{1, \frac{\rho(\mathbf{z}')}{\rho(\mathbf{z})}\},$$

and the ratio between the two density values corresponding to $\mathbf{z}$ and $\mathbf{z}'$ is equivalent to

$$H(\mathbf{z}) - H(\mathbf{z}') = 0.$$

This means that if the Hamiltonian dynamics is integrated exactly, the acceptance probability is always 1.

The numerical integrator can be constructed by using a splitting method. Specifically, the equations of motion Eq. (2.1) can be split as

$$\mathrm{d}\mathbf{z} = \mathbf{v}_1 \mathrm{d}t + \mathbf{v}_2 \mathrm{d}t, \tag{2.7}$$

and the flow map constructed separately for $\mathbf{v}_1$ and $\mathbf{v}_2$, obtaining

$$\Psi_t = \Phi_t^{\mathbf{v}_2} \circ \Phi_t^{\mathbf{v}_1}. \tag{2.8}$$

In the Hamiltonian system, for example, let

$$\mathbf{v}_1 = [\mathbf{p}^\mathsf{T}, \mathbf{0}^\mathsf{T}]^\mathsf{T}, \quad \mathbf{v}_2 = [\mathbf{0}^\mathsf{T}, \mathbf{f}^\mathsf{T}]^\mathsf{T}.$$

A general splitting method is

$$\Psi_t = \Phi^{\mathbf{v}_2}_{b_n t} \circ \Phi^{\mathbf{v}_1}_{a_n t} \circ \cdots \circ \Phi^{\mathbf{v}_2}_{b_1 t} \circ \Phi^{\mathbf{v}_1}_{a_1 t}.$$

The coefficients satisfy $\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} b_i = 1$.

The leapfrog method is a symmetric splitting:

$$\Psi_t = \Phi^{\mathbf{v}_2}_{t/2} \circ \Phi^{\mathbf{v}_1}_{t} \circ \Phi^{\mathbf{v}_2}_{t/2}.$$

For HMC, the explicit form of the leapfrog integrator is

$$\begin{aligned} \mathbf{p}_{1/2} &= \mathbf{p}_0 + \frac{1}{2}\mathbf{f}(\theta_0)\Delta t, \\ \theta_1 &= \theta_0 + \mathbf{p}_{1/2}\Delta t, \\ \mathbf{p}_1 &= \mathbf{p}_{1/2} + \frac{1}{2}\mathbf{f}(\theta_1)\Delta t, \end{aligned} \tag{2.9}$$

where Δt is the discretization step size. It can be seen that the leapfrog method is also reversible and volume preserving. Therefore, the stationarity of the HMC propagator is guaranteed [12]. The global error of the leapfrog method is $\mathcal{O}((\Delta t)^2)$ [14]. So the actual acceptance probability depends on the discretization step size Δt and the number of steps taken.

2.2 The General Theory

The HMC method produces the desired distribution, because the numerical integrator of the Hamiltonian dynamics is reversible and volume preserving [12]. Reversibility and volume preservation imply that the transition density for $\mathbf{z}$ and $\mathbf{z}'$ is symmetric, so the Metropolis criterion defined in terms of the joint density itself (Eq. (1.2)) is sufficient for guaranteeing stationarity.

When considering more general dynamics and corresponding numerical integrators with less restrictions, for example, no volume preservation, the Metropolis criterion for the ordinary HMC is no longer valid. This is because that the change of volume renders the transition probability asymmetric. Therefore, a new criterion is needed

for justifying the use of more general dynamics. The following theorem states the formal result.

Theorem 2.2.1 *Let* $\mathbf{z}$ *be the extended space variable. Define* $\mathbf{z}' = \Psi_\tau(\mathbf{z})$.

If the following conditions are satisfied:

1. Ψ_τ *is reversible under* R, *i.e.,* $\Psi_\tau^{-1}(\mathbf{z}) = R(\Psi_\tau(R(\mathbf{z})))$.

2. R *is an involution, i.e.,* $R \circ R = \mathrm{id}$,

3. *the proposed sample* $R(\mathbf{z}')$ *is accepted with probability*

$$\min\{1, \frac{\rho(R(\mathbf{z}'))}{\rho(\mathbf{z})} |\det(\partial_{\mathbf{z}} R \circ \Psi_\tau(\mathbf{z}))|\}, \tag{2.10}$$

where $\partial_{\mathbf{z}} R \circ \Psi_\tau(\mathbf{z})$ *is the Jacobian matrix. Otherwise, keep* $\mathbf{z}$,

Then $\rho(\mathbf{z})$ *is the stationary density of the Markov chain generated by* Ψ_τ *and the acceptance test* (2.10).

Proof It suffices to show detailed balance. Let $\mathbf{z}'' = R(\mathbf{z}')$, and $\rho_{\mathrm{t}}(\mathbf{z}''|\mathbf{z})$ denote the transition density. To find a computable expression for the quotient in the Metropolis-Hastings test, i.e., the relationship between $\rho_{\mathrm{t}}(R(\mathbf{z}')|\mathbf{z})$ and $\rho_{\mathrm{t}}(\mathbf{z}|R(\mathbf{z}'))$,write

$$\begin{aligned}
&\rho_{\mathrm{t}}(\mathbf{z}|\mathbf{z}'') \\
&= \delta(\mathbf{z} - R(\Psi_\tau(R(\mathbf{z}')))) \\
&= \delta(R(\Psi_\tau(\mathbf{z})) - R(\Psi_\tau(R(\Psi_\tau(R(\mathbf{z}')))))) |\det(\partial_{\mathbf{z}} R \circ \Psi_\tau(\mathbf{z}))| \\
&= \delta(R(\Psi_\tau(\Psi_\tau^{-1}(\mathbf{z}'))) - R(\Psi_\tau(\mathbf{z}))) |\det(\partial_{\mathbf{z}} R \circ \Psi_\tau(\mathbf{z}))| \\
&= \delta(R(\mathbf{z}') - R(\Psi_\tau(\mathbf{z}))) |\det(\partial_{\mathbf{z}} R \circ \Psi_\tau(\mathbf{z}))| \\
&= \rho_{\mathrm{t}}(\mathbf{z}''|\mathbf{z}) |\det(\partial_{\mathbf{z}} R \circ \Psi_\tau(\mathbf{z}))|.
\end{aligned}$$

The second equality holds because for a function $\mathbf{g}$ compositing with the delta function, $\delta(\mathbf{z}) = \delta(\mathbf{g}(\mathbf{z}))|\det(\partial_{\mathbf{z}}\mathbf{g}(\mathbf{z}))|$. The Metropolis-Hastings criterion needed to ensure the detailed balance (Eq. (2.3)) is

$$\begin{aligned}
&\min\{1, \frac{\rho(R(\mathbf{z}'))\rho_{\mathrm{t}}(\mathbf{z}|R(\mathbf{z}'))}{\rho(\mathbf{z})\rho_{\mathrm{t}}(R(\mathbf{z}')|\mathbf{z})}\} \\
&= \min\{1, \frac{\rho(R(\mathbf{z}'))\rho_{\mathrm{t}}(R(\mathbf{z}')|\mathbf{z})}{\rho(\mathbf{z})\rho_{\mathrm{t}}(R(\mathbf{z}')|\mathbf{z})}\}|\det(\partial_{\mathbf{z}}R \circ \Psi_\tau(\mathbf{z}))| \\
&= \min\{1, \frac{\rho(R(\mathbf{z}'))}{\rho(\mathbf{z})}|\det(\partial_{\mathbf{z}}R \circ \Psi_\tau(\mathbf{z}))|\}.
\end{aligned}$$

□

Note that theorem 2.2.1 is not the entire MCMC method, but one key substep. The reason for applying R to $\mathbf{z}'$ is to fit this substep into the Metropolis-Hasting framework. The theorem has a general form for R. In ordinary HMC, Ψ_τ is reversible under $R : [\theta^\mathsf{T}, \mathbf{p}^\mathsf{T}]^\mathsf{T} \mapsto [\theta^\mathsf{T}, -\mathbf{p}^\mathsf{T}]^\mathsf{T}$. So R is volume preserving and satisfies $\rho(R(\mathbf{z})) = \rho(\mathbf{z})$. It is found that there is no obvious benefit for R not being volume preserving and satisfying $\rho(R(\mathbf{z})) = \rho(\mathbf{z})$. With these restrictions, the Metropolis criterion (2.10) is reduced to

$$\min\{1, \frac{\rho(\mathbf{z}')}{\rho(\mathbf{z})}|\det(\partial_{\mathbf{z}}\Psi_\tau(\mathbf{z}))|\}. \tag{2.11}$$

Moreover, Thm. 2.2.1 states nothing about ergodicity, In ordinary HMC, the momenta $\mathbf{p}$ are randomized based on the distribution of $\mathbf{p}$ to ensure ergodicity. For a generalized HMC, a randomization of $\mathbf{z}$ is also necessary, and the randomization should preserve the density of $\mathbf{z}$ to keep stationarity. Since the density of the state space variable θ is only known up to a normalizing constant, it is more convenient to only randomize the auxiliary variables in $\mathbf{z}$, for which the density can be designed for it to be easy to generate independent random variables. For example, the momenta $\mathbf{p}$ in the ordinary HMC has a Gaussian distribution.

Recall that in ordinary HMC, the acceptance probability is always 1 if the dynamical system is integrated exactly. Now the question is whether this condition is also satisfied by the generalized acceptance test. The answer is yes if the dynamics is designed based on the continuity equation (Eq. (2.5)). To see this, consider one

Monte Carlo step, where the starting position $\mathbf{z}$ is given and $\mathbf{z}' = \Phi_t(\mathbf{z})$ with Φ_t being the flow of the dynamics. It is straightforward to show that for any t, the ratio in (2.11) is always 1, if $\mathbf{v}$ and ρ satisfy the continuity equation, and the dynamics is integrated exactly.

In practice, the dynamical system may be solved only by a numerical integrator. In order to guarantee stationarity (see Thm. 2.2.1), the numerical integrator must be reversible. In addition, the discretization error introduced by the numerical integrator makes the actual acceptance probability less than 1. So the efficiency of the generalized HMC method also depends on the accuracy of the numerical integrator.

2.2.1 The Meta-algorithm

Based on the analysis above, the framework for generalized HMC, called the "meta-algorithm" is described as follow:

1. Construct a dynamical system, of which the joint density $\rho(\mathbf{z})$ and the vector field $\mathbf{v}(\mathbf{z})$ satisfy the continuity equation (Eq. (2.5)). The dynamical system should be reversible under R. The joint density should marginalize to the target density $\rho(\theta)$, and the marginal density of auxiliary variables should be easy to produce independent random samples.

2. Find a numerical integrator for the dynamical system. The numerical integrator should also be reversible. Find the Jacobian of the map of the numerical integrator.

3. Let Ψ_τ be the map of the numerical integrator for time interval τ, and the map R be volume preserving and satisfy $\rho(R(\mathbf{z})) = \rho(\mathbf{z})$. A new sample θ' is obtained from the following algorithm:

   (a) Let θ be given.

   (b) Randomize the auxiliary variable according to the marginal distribution of the auxiliary variable to form $\mathbf{z}$.

(c) Compute $\mathbf{z}' = \Psi_\tau(\mathbf{z})$,

(d) Accept θ', which is contained in $\mathbf{z}'$, with probability

$$\min\{1, \frac{\rho(\mathbf{z}')}{\rho(\mathbf{z})} |\det(\partial_{\mathbf{z}} \Psi_\tau(\mathbf{z}))|\},$$

otherwise keep θ.

Note that in this algorithm, R becomes irrelevant if the auxiliary variable is randomized in each step.

2.3 Examples

This section illustrates the use of the "meta-algorithm" proposed in the last section.

2.3.1 Example 1: Generalized Hamiltonian System

This toy example illustrates the procedure of designing a new dynamical system with some desired properties by using the continuity equation (2.6). In ordinary HMC, the probability density of $\mathbf{p}$ is Gaussian. One can also choose other densities for $\mathbf{p}$, such as the Student's t distribution, which is known to have heavy tail and may be useful [18]. The goal is to find the equations of motion given the target density.

Let $\mathbf{p}$ follow the Student's t distribution:

$$\rho_{\mathbf{p}}(\mathbf{p}) \propto (1 + \frac{\mathbf{p}^\mathsf{T}\mathbf{p}}{\nu})^{(-\frac{N+\nu}{2})},$$

where N is the dimension of $\mathbf{p}$ and ν is the degree of freedom of the Student's t distribution. Given the target density $\rho_\theta(\theta) \propto \exp(-U(\theta))$, try $\rho(\theta, \mathbf{p}) \propto \exp(-H(\theta, \mathbf{p}))$ with $H(\theta, \mathbf{p}) = U(\theta) + K(\mathbf{p})$ where

$$K(\mathbf{p}) = \frac{N+\nu}{2} \ln(1 + \frac{\mathbf{p}^\mathsf{T}\mathbf{p}}{\nu}).$$

Assuming volume preservation, then

$$\nabla_\theta H \cdot \mathbf{v}_\theta + \nabla_{\mathbf{p}} H \cdot \mathbf{v}_{\mathbf{p}} = 0,$$

namely,

$$\mathbf{f} \cdot \mathbf{v}_\theta = \frac{\nu + N}{\nu + \mathbf{p}^\mathsf{T}\mathbf{p}} \mathbf{p} \cdot \mathbf{v}_\mathbf{p}.$$

A divergence free system can be easily found:

$$\begin{aligned} \mathrm{d}\theta &= \mathbf{v}_\theta \mathrm{d}t = \frac{\nu + N}{\nu + \mathbf{p}^\mathsf{T}\mathbf{p}} \mathbf{p} \mathrm{d}t, \\ \mathrm{d}\mathbf{p} &= \mathbf{v}_\mathbf{p} \mathrm{d}t = \mathbf{f} \mathrm{d}t. \end{aligned}$$

It can be seen that when $\nu \to \infty$, this dynamical system converges to the ordinary Hamiltonian system. This is expected, since in the ordinary HMC, the probability density for $\mathbf{p}$ is Gaussian and Student's t $\to$ Gaussian as $\nu \to \infty$.

2.3.2 Example 2: Nosé-Hoover Thermostat

The extended space may contain auxiliary variables in addition to the momenta $\mathbf{p}$. For example, the Nosé-Hoover thermostat employs the extended system

$$\begin{aligned} \mathrm{d}\theta &= \mathbf{p} \mathrm{d}t, \\ \mathrm{d}\mathbf{p} &= \mathbf{f} - \frac{\xi}{\mu} \mathbf{p} \mathrm{d}t, \\ \mathrm{d}\xi &= (\mathbf{p}^\mathsf{T}\mathbf{p} - N) \mathrm{d}t, \end{aligned} \tag{2.12}$$

where $\mu > 0$ is "thermal mass". Let $\mathbf{z} = [\theta^\mathsf{T}, \mathbf{p}^\mathsf{T}, \xi]^\mathsf{T}$ be the extended space variable. The Nosé-Hoover thermostat has the invariant density

$$\rho(\mathbf{z}) \propto \exp(-H(\mathbf{z})), \quad H(\mathbf{z}) = U(\theta) + \frac{1}{2}\mathbf{p}^\mathsf{T}\mathbf{p} + \frac{1}{2\mu}\xi^2.$$

The dynamical system has non-zero compressibility

$$\nabla \cdot \mathbf{v} = -\frac{\xi}{\mu} N.$$

Leimkuhler and Reich [19] develop a generalized HMC method based on the Nosé-Hoover thermostat .

2.3.3 Example 3: Isokinetic Ensemble

In this example, an MCMC sampler using the dynamics of the isokinetic ensemble is constructed by the meta-algorithm. The isokinetic ensemble [20] maintains the kinetic energy of the system constant. Shown is the process of using the meta-algorithm—from designing the dynamical system to computing the Jacobian. The dynamical system derived here is equivalent to the dynamics of isokinetic ensemble in [21]. Numerical experiments are performed to justify the new sampler.

First, by using the continuity equation, find the dynamical system which maintains the kinetic energy constant. The system is not necessarily volume preserving. To find such system, following [22], define H as a function of $\mathbf{z}$ satisfying

$$\nabla H \cdot \mathbf{v} = 0.$$

The equation above makes H a conserved quantity of the dynamics. Note that here H is not necessarily the negative logarithm of the density plus a constant. Write the density as

$$\rho(\mathbf{z}) \propto \exp(-\omega(\mathbf{z}))\zeta(H(\mathbf{z})),$$

where ζ is any function of H (e.g., $\exp(-H)$), and $\omega(\mathbf{z})$ is called the *compressibility integral*, because if assuming $\nabla H \cdot \mathbf{v} = 0$, ω satisfies

$$\nabla\omega \cdot \mathbf{v} = \nabla \cdot \mathbf{v}. \tag{2.13}$$

Note that the existence of a solution of Eq. (2.13) is possible only in special cases.

Let $K = \mathbf{p}^\mathsf{T}\mathbf{p}/2$ be the kinetic energy. The goal is to find a dynamical system that conserves K. Therefore, choose $H = K$. The marginal probability density of $\mathbf{p}$ needs to be a function of K, e.g., Gaussian: $\rho(\mathbf{p}) \propto \exp(-\mathbf{p}^\mathsf{T}\mathbf{p}/2)$, Student's t: $\rho(\mathbf{p}) \propto (1 + \mathbf{p}^\mathsf{T}\mathbf{p}/\nu)^{-(\nu+N)/2}$, or a delta function: $\rho(\mathbf{p}) \propto \delta(\mathbf{p}^\mathsf{T}\mathbf{p} - E)$ where E is a constant. The function ζ allows a flexible choice on the density of $\mathbf{p}$. Note that to allow the delta function to be considered as a choice of ζ, δ may be taken as a mollified delta function as in [11].

Because $\exp(-\omega(\mathbf{z}))\zeta(K(\mathbf{p}))$ must marginalize to $\rho(\theta)$, try

$$\exp(-\omega(\theta)) = \exp(-U(\theta)) \propto \rho(\theta),$$

and obtain $\omega = U$. The equations of motion $\mathbf{v}_\theta$ and $\mathbf{v}_\mathbf{p}$ must satisfy

$$\begin{aligned} &\nabla K \cdot \mathbf{v} = \mathbf{p} \cdot \mathbf{v}_\mathbf{p} = 0, \\ &\nabla \cdot \mathbf{v} = \nabla_\theta \cdot \mathbf{v}_\theta + \nabla_\mathbf{p} \cdot \mathbf{v}_\mathbf{p} = -\mathbf{v}_\theta \cdot \mathbf{f} = \nabla\omega \cdot \mathbf{v}. \end{aligned} \tag{2.14}$$

A solution $\mathbf{v}$ of the system of equations above gives a desired dynamical system.

For the first equation, the vector $\mathbf{v}_\mathbf{p}$ must be orthogonal to $\mathbf{p}$. It is reasonable to make $\mathbf{v}_\mathbf{p}$ depend on $\mathbf{f}$, and be as close to $\mathbf{f}$ as possible. In particular, choose $\mathbf{v}_\mathbf{p}$ to minimize

$$\|\mathbf{f} - \mathbf{v}_\mathbf{p}\|_2, \quad \text{s.t. } \mathbf{p} \cdot \mathbf{v}_\mathbf{p} = 0.$$

By using a Lagrange multiplier, it is easy to see that the optimal $\mathbf{v}_\mathbf{p}$ is

$$\mathbf{v}_\mathbf{p} = (\mathbf{I} - \frac{\mathbf{p}\mathbf{p}^\mathsf{T}}{\mathbf{p}^\mathsf{T}\mathbf{p}})\mathbf{f}.$$

Then $\mathbf{v}_\theta$ can be solved by plugging $\mathbf{v}_\mathbf{p}$ into the second equation of (2.14):

$$\nabla_\theta \cdot \mathbf{v}_\theta + (1 - N)\frac{\mathbf{f}^\mathsf{T}\mathbf{p}}{\mathbf{p}^\mathsf{T}\mathbf{p}} = -\mathbf{v}_\theta \cdot \mathbf{f}.$$

So one solution can be obtained by observation:

$$\mathbf{v}_\theta = \frac{N-1}{\mathbf{p}^\mathsf{T}\mathbf{p}}\mathbf{p}.$$

The final equations of motion are

$$\begin{aligned} \mathrm{d}\theta &= \mathbf{v}_\theta \mathrm{d}t = \frac{N-1}{\mathbf{p}^\mathsf{T}\mathbf{p}}\mathbf{p}\mathrm{d}t, \\ \mathrm{d}\mathbf{p} &= \mathbf{v}_\mathbf{p}\mathrm{d}t = \mathbf{f}\mathrm{d}t - \frac{\mathbf{p}^\mathsf{T}\mathbf{f}}{\mathbf{p}^\mathsf{T}\mathbf{p}}\mathbf{p}\mathrm{d}t. \end{aligned}$$

Note that when $\rho(\mathbf{p}) \propto \delta(\mathbf{p}^\mathsf{T}\mathbf{p} - N)$, the equations of motion above are equivalent to those of the isokinetic ensemble in [21]:

$$\begin{aligned} \mathrm{d}\theta &= \frac{N-1}{N}\mathbf{p}\mathrm{d}t, \\ \mathrm{d}\mathbf{p} &= \mathbf{f}\mathrm{d}t - \frac{\mathbf{p}^\mathsf{T}\mathbf{f}}{\mathbf{p}^\mathsf{T}\mathbf{p}}\mathbf{p}\mathrm{d}t. \end{aligned}$$

with invariant density

$$\rho(\theta, \mathbf{p}) \propto \exp(-\frac{\mathbf{p}^\mathsf{T}\mathbf{p}}{N}U(\theta))\delta(\mathbf{p}^\mathsf{T}\mathbf{p} - N).$$

It is easy to see that this dynamical system conserves the kinetic energy, since the vector $\mathbf{v_p}$ is perpendicular to $\mathbf{p}$ (see Eq. (2.14)).

A time reversible and kinetic energy conserved integrator [23] can be obtained by the operator splitting scheme (Strang splitting):

$$\Psi_{\Delta t} = \Phi^{\mathbf{v}_2}_{\Delta t/2} \circ \Phi^{\mathbf{v}_1}_{\Delta t} \circ \Phi^{\mathbf{v}_2}_{\Delta t/2},$$

with $\mathbf{v}_1 = [((N-1)/N)\mathbf{p}^\mathsf{T}, \mathbf{0}^\mathsf{T}]^\mathsf{T}$ and $\mathbf{v}_2 = [\mathbf{0}^\mathsf{T}, \mathbf{f}^\mathsf{T} - (\mathbf{p}^\mathsf{T}\mathbf{f}/\mathbf{p}^\mathsf{T}\mathbf{p})\mathbf{p}^\mathsf{T}]^\mathsf{T}$.

The following are the substeps of the $\Psi_{\Delta t}$ mapping as given in [23][Sec. IV.B]. Let the discretization step size be Δt. Let $\dot{c}$ denote the time derivative of a function $c(t)$.

Step 1: $\Phi^{\mathbf{v}_2}_{\Delta t/2}$ ("kick"): Evaluate $s(\Delta t/2)$ and $\dot{s}(\Delta t/2)$, where

$$s(t) = \frac{a}{b}(\cosh(t\sqrt{b}) - 1) + \frac{1}{\sqrt{b}}\sinh(t\sqrt{b}),$$

and

$$a = \frac{\mathbf{p}^\mathsf{T}\mathbf{f}}{\mathbf{p}^\mathsf{T}\mathbf{p}}, \quad b = \frac{\mathbf{f}^\mathsf{T}\mathbf{f}}{\mathbf{p}^\mathsf{T}\mathbf{p}}.$$

Then update $\mathbf{p}$:

$$\mathbf{p}_{1/2} = \frac{\mathbf{p}_0 + s(\Delta t/2)\mathbf{f}_0}{\dot{s}(\Delta t/2)}.$$

Step 2: $\Phi^{\mathbf{v}_1}_{\Delta t}$ ("drift"): Update θ

$$\theta_1 = \theta_0 + \frac{N-1}{N}\mathbf{p}_{1/2}\Delta t,$$

and then calculate the new force.

Step 3: $\Phi^{\mathbf{v}_2}_{\Delta t/2}$ ("kick"): Evaluate $s(\Delta t/2)$ and $\dot{s}(\Delta t/2)$ and then update $\mathbf{p}$:

$$\mathbf{p}_1 = \frac{\mathbf{p}_{1/2} + s(\Delta t/2)\mathbf{f}_1}{\dot{s}(\Delta t/2)}.$$

The Jacobian of $\Psi_{\Delta t}$ can be obtained by combining the Jacobians of $\Phi^{\mathbf{v}_2}_{\Delta t/2}$ and $\Phi^{\mathbf{v}_1}_{\Delta t}$. For a general m-step splitting method

$$\Psi_{\Delta t} = \Phi^{\mathbf{v}_1}_{a_m \Delta t} \circ \Phi^{\mathbf{v}_2}_{b_m \Delta t} \circ \cdots \circ \Phi^{\mathbf{v}_1}_{a_1 \Delta t} \circ \Phi^{\mathbf{v}_2}_{b_1 \Delta t},$$

the Jacobian of the entire map $\Psi_{\Delta t}$ can be obtained by

$$J = \prod_{i=1}^{m} J_i,$$

where J_i is the Jacobian in step i.

The Jacobian of $\Phi^{\mathbf{v}_2}_{\Delta t/2}$ in the "Kick" part can be obtained by integrating the divergence. A similar derivation can be found in [11]. Specifically, the Jacobian of $\Phi^{\mathbf{v}_2}_t$ for $0 \leq t \leq \Delta t/2$ satisfies:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \log J(t) &= \nabla \cdot \mathbf{v}_2(\mathbf{z}), \\ \log J(0) &= 0. \end{aligned}$$

Let $\dot{r}(t) = \mathbf{f}^\mathsf{T}\mathbf{p}/(\mathbf{p}^\mathsf{T}\mathbf{p})$. When $\mathbf{f}$ is fixed, $s(t)$ satisfies

$$\ddot{s}(t) = \dot{s}(t)\dot{r}(t).$$

So $\dot{s}(t) = \exp(r(t))$. And $\nabla \cdot \dot{\mathbf{z}} = (1-N)\dot{r}(t)$, so

$$\log \dot{s}(t) = r(t) = \log J(t)/(1-N).$$

Therefore,

$$J(t) = \dot{s}(t)^{-(N-1)}.$$

The Jacobian of the "Drift" part is 1.

This numerical integrator is reversible under $R : [\theta^\mathsf{T}, \mathbf{p}^\mathsf{T}]^\mathsf{T} \mapsto [\theta^\mathsf{T}, -\mathbf{p}^\mathsf{T}]^\mathsf{T}$ and conserves the kinetic energy.

Now all components of the MCMC sampler, a dynamical system with desired properties satisfying the continuity equation, a reversible numerical integrator, and the Jacobian of the integrator, are obtained. Samples can be produced by using the third step of the meta-algorithm (Sec. 2.2.1).

Table 2.1.
Effective sample size per 1000 force evaluations for Hamiltonian HMC

$\tau\backslash\nu$	6	8	10	12
4	1.44	2.51	3.03	2.55
5	2.04	4.41	4.13	3.65
6	-	1.52	3.42	3.19

A test problem is designed to run both ordinary HMC and isokinetic HMC for a comparison. The problem is a high-dimensional mixture of two Gaussians. Specifically, the first dimension is the mixture of two Gaussians located at ± 2.5 with standard deviation 1. Another 128 dimensions are all Gaussians located at the origin with standard deviations uniformly distributed from 1 to 2. Note that all dimensions of θ are independent from each other. Though simple, the test problem is high dimensional and multimodal, and it is not sensitive to the choice of the integration duration in one MC step, an important parameter for the MCMC method. Choose the function of interest to be a sigmoid function in the first dimension. As is known [24], the sigmoid function is similar, in shape, to the eigenfunction corresponding to the subdominant eigenvalue of the propagator for a double well potential problem. The discretization step size $\Delta t = 0.5$, the number of integration steps is 10, and the number of samples is 10^6.

Tables 2.1 and 2.2 show the effective sample sizes per 1000 force evaluations for each method, for varying values of the duration τ of an MC step and the number of integrator steps ν per MC step. A pair of parameter values that maximizes the number of effective samples per integrator step is in the center of each table. A dash indicates a failure of all proposed moves. It can be seen that the isokinetic method performs slightly better than the Hamiltonian method and is less sensitive to tuning parameters.

Table 2.2.
Effective sample size per 1000 force evaluations for isokinetic HMC

$\tau\backslash\nu$	6	8	10	12
4	3.16	3.52	2.81	3.20
5	3.83	4.52	4.91	4.84
6	0.72	4.11	3.60	3.29

2.3.4 Example 4: Variable Mass Methods

The variable mass HMC is proposed by Girolami and Calderhead [18] for Bayesian statistics. The expected Fisher information matrix is used to form the mass matrix embedded in the kinetic energy. This method samples the anisotropic basin of the potential energy more efficiently than the ordinary HMC. The integrator is symplectic, but requires solving a nonlinear equation involving the mass matrix $\mathbf{M}(\theta)$.

The choice of the mass matrix can also be targeted for reducing the energy barrier of the potential energy function, hence make sampling multimodal problems easier. In addition, by removing the volume-preserving constraint of the dynamics, the integrator can be made explicit to avoid solving the nonlinear equation.

The dynamical system of the variable mass HMC has invariant density $\rho(\theta, \mathbf{p}) \propto \exp(-H(\theta, \mathbf{p}))$ with the Hamiltonian

$$H(\theta, \mathbf{p}) = U(\theta) + \frac{1}{2}\mathbf{p}^{\mathsf{T}}\mathbf{M}^{-1}(\theta)\mathbf{p} + \frac{1}{2}\log\det\mathbf{M}(\theta).$$

The logarithm determinant term comes from the normalizing constant of the Gaussian distribution with covariance matrix $\mathbf{M}$. Note that the marginal distribution of θ is still $\propto \exp(-U(\theta))$.

The equations of motion of the variable mass Hamiltonian system is

$$\begin{aligned}
\mathrm{d}\theta &= \mathbf{M}^{-1}\mathbf{p}\mathrm{d}t, \\
\mathrm{d}\mathbf{p} &= \mathbf{f}\mathrm{d}t - \frac{1}{2}\sum_{k=1}^{N}(\mathrm{tr}(\mathbf{M}^{-1}\mathbf{M}_k) - \mathbf{p}^{\mathsf{T}}\mathbf{M}^{-1}\mathbf{M}_k\mathbf{M}^{-1}\mathbf{p})\mathbf{e}_k\mathrm{d}t.
\end{aligned}$$

where

$$\mathbf{M}_k = \frac{\partial\mathbf{M}(\theta)}{\partial\theta_k},$$

and $\mathbf{e}_k$ is the unit vector of the kth coordinate direction.

To find explicit integrators, change $(\theta, \mathbf{p})$ to $(\theta, \mathbf{M}(\theta)\mathbf{v})$. Note that here the symbol $\mathbf{v}$ is reused to denote the velocity. Then the Hamiltonian is changed to

$$H(\theta, \mathbf{v}) = U(\theta) + \frac{1}{2}\mathbf{v}^{\mathsf{T}}\mathbf{M}\mathbf{v} - \frac{1}{2}\log\det\mathbf{M}(\theta), \tag{2.15}$$

and the equations of motion become

$$\begin{aligned} &\mathrm{d}\theta = \mathbf{v}\mathrm{d}t, \\ &\mathbf{M}\mathrm{d}\mathbf{v} = \mathbf{f}(\theta)\mathrm{d}t + \sum_{k=1}^{N} \frac{1}{2}(\mathrm{tr}(\mathbf{M}_k\mathbf{M}^{-1}) - \mathbf{v}^{\mathsf{T}}\mathbf{M}_k\mathbf{v} - \sum_{k=1}^{N} \mathbf{M}_k\mathbf{M}^{-1})\mathbf{e}_k\mathrm{d}t. \end{aligned} \quad (2.16)$$

This can be solved by explicit integrators. Consider quadratic ODE system

$$\mathrm{d}v_i = (\frac{1}{2}\mathbf{v}^{\mathsf{T}}\mathbf{A}_i\mathbf{v} + \mathbf{b}_i^{\mathsf{T}}\mathbf{v} + \gamma_i)\mathrm{d}t,$$

where, without loss of generality, $\mathbf{A}_i$ is symmetric. A time-symmetric discretization, which is merely linearly implicit, is

$$\frac{v_i^1 - v_i^0}{\Delta t} = \frac{1}{2}(\mathbf{v}^0)^{\mathsf{T}}\mathbf{A}_i\mathbf{v}^1 + \frac{1}{2}\mathbf{b}_i^{\mathsf{T}}(\mathbf{v}^0 + \mathbf{v}^1) + \gamma_i.$$

Applying this to Eq. (2.16), we have

$$(\mathbf{I} + \frac{\Delta t}{2}\sum_{k=1}^{N} \mathbf{M}^{-1}\mathbf{e}_k(\mathbf{v}^0)^{\mathsf{T}}\mathbf{M}_k)\mathbf{v}^1 = \mathbf{v}^0 - \Delta t\mathbf{M}^{-1}(\nabla_\theta U^+ + \sum_{k=1}^{N} \mathbf{M}_k\mathbf{M}^{-1}\mathbf{e}_k),$$

where

$$U^+(\theta) = U(\theta) - \log\det(\mathbf{M}(\theta))/2. \quad (2.17)$$

The Jacobian of the map of the integrator is

$$J = \frac{\det(\mathbf{I} - (\Delta t/2)\sum_k \mathbf{M}^{-1}\mathbf{e}_k(\mathbf{v}^1)^{\mathsf{T}}\mathbf{M}_k)}{\det(\mathbf{I} + (\Delta t/2)\sum_k \mathbf{M}^{-1}\mathbf{e}_k(\mathbf{v}^0)^T\mathbf{M}_k)}.$$

By using the variable mass, the potential energy is changed to U^+ with an extra term added (Eqs. (2.15) and (2.17)). Therefore, by choosing proper $\mathbf{M}(\theta)$, some useful properties can be brought into the dyanmics, such as reducing the energy barrier of the potential energy. For example, choose $\mathbf{M}$ to be

$$\mathbf{M} = \frac{1}{\mu}(\mathbf{I} - (1 - \mu^N)\mathbf{m}\mathbf{m}^T),$$

where μ is

$$\mu = 1 - (1 - \mu_0)\exp(-\sigma\|\theta - \theta_0\|^2)),$$

$\mathbf{m}$ is the direction of the barrier and $\mathbf{m}^{\mathsf{T}}\mathbf{m} = 1$, θ_0 is the location of the barrier, σ is a parameter that controls the speed of the change of the mass, and μ_0 determines the

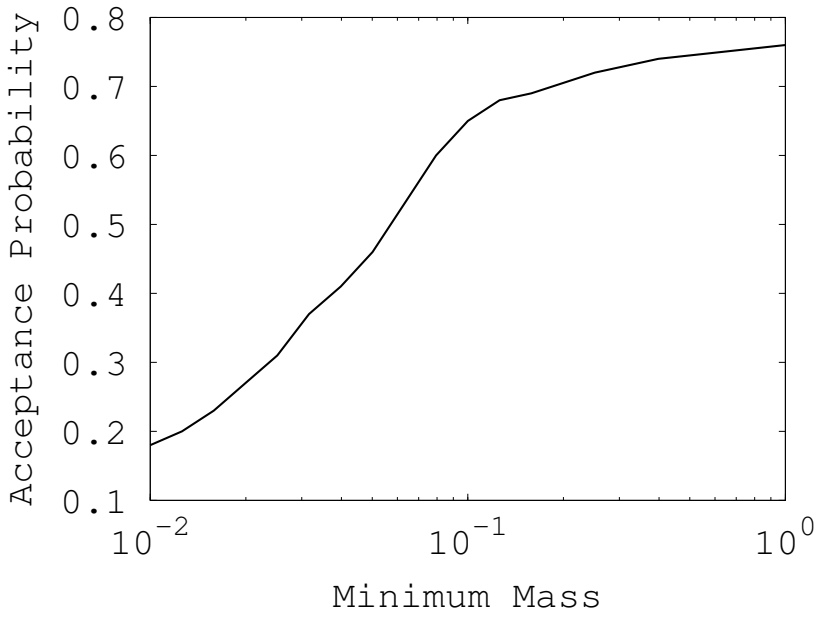


Figure 2.1. Acceptance probability vs. the minimum mass for the variable mass method in the double-well potential problem.

minimum of the mass. The idea here is to increase the speed of particles to across the region of the barrier, by decreasing the mass. The closer the particle to the saddle point, the higher the speed. And in the regions far away from the barrier, the mass remains constant. Note that to use this method, prior knowledge of the barrier is required.

This method is tested on the same problem as in Sec. 2.3.3. The autocorrelation time, computed by `Acor`, gives the ratio of the sample size to the effective sample size. Results are displayed in Figs 2.1 and 2.2 for $\mathbf{m}$ in the direction of θ_1, $\theta_0 = \mathbf{0}$, $\sigma = 1$, and a range of values of the minimum mass. For a fairly wide range of values of the minimum mass, the variable mass method produces about 100% more effective samples.

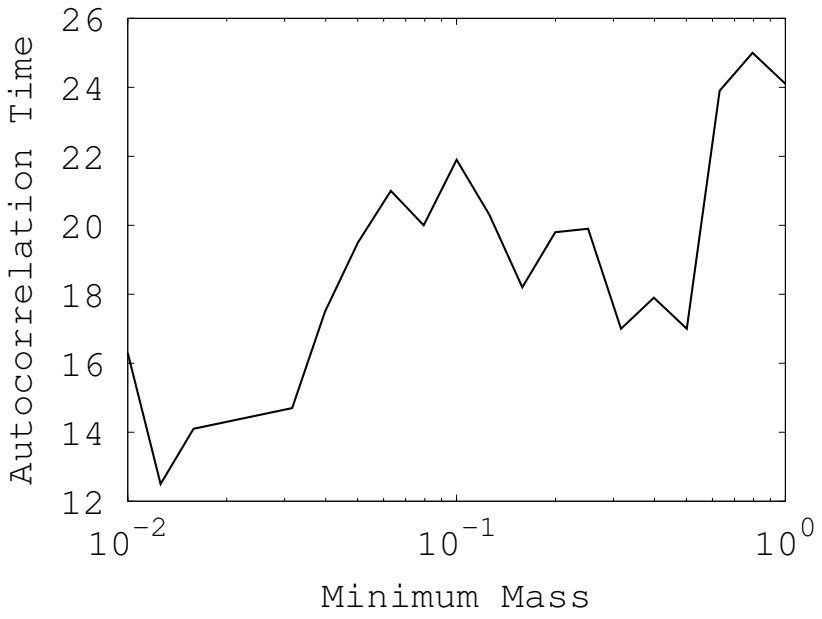


Figure 2.2. Autocorrelation time vs. the minimum mass for the variable mass method in the double-well potential problem.

2.4 Discussion and Conclusion

More possibilities of designing efficient dynamics based MCMC methods are opened up by a more general framework—the meta-algorithm proposed in this chapter. The abilities such as conquering energy barriers are benefits that brought by the general framework. However, the example methods developed here still have limitations, due to the lack of knowledge about the location of the barriers. Better methods are yet to be discovered by combining other techniques, such as adaptively locating the barriers and parallel tempering.

3 STOCHASTIC GRADIENT SAMPLERS WITH REDUCED BIAS

The computational cost of dynamics-based samplers is primarily that of the evaluation of the gradient of the potential energy function. In many applications, such as machine learning problems with tremendous data, the evaluation of gradients is too expensive for some of the models to be feasible in practice. For instance, consider probabilistic models of which the potential energy function consists of the summation of n negative log likelihood functions, where n is the number of data examples. Nowadays, most important machine learning applications rely on the use of extremely large datasets for training. This usually results in unbearably expensive gradient computations.

To make probabilistic models, especially Bayesian models, useful in machine learning applications, the idea of stochastic gradients was adopted for sampling methods based on stochastic differential equations such as Brownian dynamics [4] and Langevin dynamics [6]. A stochastic gradient uses the gradient obtained from a random subset of the data to approximate the full gradient. The size of the subset is usually much smaller than the size of the entire dataset.

Methods introduced in Chap. 2, which are referred as *rigorous* sampling, combine a dynamical system with the Metropolis step. However, the Metropolis step is incompatible with the idea of stochastic gradient, because the evaluation of the potential energy function requires using the entire dataset which cancels the benefit of using stochastic gradients. In some molecular dynamics applications, the Metropolis step is indeed omitted, at least for certain types of samplers such as MALA [25], for there being theoretical justification regarding the existence of the modified density and the controllability of the error [13]. For the same reason, sampling methods based on the discretization of stochastic differential equations are also widely used in practical machine learning problems.

Previous attempts for adopting stochastic gradients in dynamics-based samplers are comparatively half-hearted—the full gradient is simply replaced by the stochastic gradient. However, stochastic gradients introduce an amount of additional noise into the dynamical system, hence change the stationary probability density. This change is undesirable and can be harmful if not controlled. To control the noise, a smaller discretization step size must be used, which offsets the benefit brought by the use of a stochastic gradient.

In this chapter, the additional noise is analyzed and a formula for estimating it adaptively is obtained. Two methods are also designed to reduce the bias introduced by the stochastic gradient. The first method is based on the estimate of the noise, and the second uses a new variable and its corresponding equation of motion to automatically remove, at least partially, the bias.

The rest of the chapter is organized as follow: Sec. 3.1 briefly reviews the related background; Sec. 3.2 presents the analysis of the noise and the first method; Sec. 3.3 presents the second method; Sec. 3.4 compares the proposed methods with previous methods on synthetic and real machine learning applications; and Sec. 3.5 concludes the chapter with a discussion.

3.1 Background

Recall Eqs. (1.3) and (1.4) for Langevin dynamics, with damping coefficient A, the stochastic differential equations are,

$$\begin{aligned} \mathrm{d}\theta &= \mathbf{p}\mathrm{d}t, \\ \mathrm{d}\mathbf{p} &= \mathbf{f}(\theta)\mathrm{d}t - A\mathbf{p}\mathrm{d}t + \sqrt{2A}\mathrm{d}\mathbf{w}, \end{aligned}$$

and for Brownian dynamics, the SDEs are

$$\mathrm{d}\theta = \mathbf{f}(\theta)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{w}.$$

By dropping the Metropolis step, SDE based methods are well suited for the adoption of stochastic gradients, since the computation requires data present only in the dynamics.

Following the definition of the stochastic gradient in Sec. 1.1.4, Eq. (1.5), let

$$\mathbf{f}_i(\theta) = -\nabla_\theta U_i(\theta), \tag{3.1}$$

whence the force can be written as

$$\mathbf{f} = \sum_{i=1}^{n} \mathbf{f}_i.$$

And the stochastic force can be written as

$$\tilde{\mathbf{f}} = \sum_{i=1}^{n} (1 + r_i)\mathbf{f}_i.$$

The SGLD algorithm [4] uses $\tilde{\mathbf{f}}(\theta)$ and Brownian dynamics integrated by the Euler-Maruyama method to generate samples:

$$\theta_1 = \theta_0 + \tilde{\mathbf{f}}(\theta_0)\Delta t + \sqrt{2\Delta t}\mathbf{a},$$

where $\mathbf{a}$ denotes N independent standard Gaussian random variables.

Similar to SGLD, SGHMC [6] uses the stochastic gradient to replace the full gradient in the ordinary Langevin dynamics. SGHMC uses a modified Langevin dynamics integrated by a simple numerical integrator,

$$\begin{aligned} \theta_1 &= \theta_0 + \mathbf{p}_0\Delta t, \\ \mathbf{p}_1 &= \tilde{\mathbf{f}}(\theta_1)\Delta t - A\mathbf{p}_0\Delta t + (2A\Delta t\mathbf{I} - 2\Delta t\hat{\mathbf{B}}(\theta_1))_{1/2}\mathbf{a}, \end{aligned} \tag{3.2}$$

where $\mathbf{M}_{1/2}$ satisfies $\mathbf{M}_{1/2}(\mathbf{M}_{1/2})^{\mathsf{T}} = \mathbf{M}$, and $\hat{\mathbf{B}}(\theta)$ is an $N \times N$ matrix intended to offset the noise in $\tilde{\mathbf{f}}$. In the actual implementation, $\hat{\mathbf{B}}(\theta)$ is simply set to zero. So

3.1.1 The Fokker-Planck Equation

Consider the following general systems of stochastic differential equations

$$\mathrm{d}\mathbf{z} = \mathbf{v}(\mathbf{z})\mathrm{d}t + (2\mathbf{D}(\mathbf{z}))_{1/2}\mathrm{d}\mathbf{w}, \tag{3.3}$$

where $\mathbf{z}$ is a general extended space variable, and $\mathbf{D}$ is an $N \times N$ matrix given that the dimension of $\mathbf{z}$ is N. Taking Langevin dynamics as an example, it has $\mathbf{z} = [\theta^\mathsf{T}, \mathbf{p}^\mathsf{T}]^\mathsf{T}$, $\mathbf{v} = [\mathbf{p}^\mathsf{T}, \mathbf{f}^\mathsf{T} - A\mathbf{p}^\mathsf{T}]^\mathsf{T}$, and $\mathbf{D} = \mathrm{diag}(\mathbf{0}, A\mathbf{I})$.

For deterministic systems, the invariant density ρ and the dynamics described by the vector field $\mathbf{v}$ must satisfy the (stationary) continuity equation (2.5). For stochastic systems, the continuity equation (2.4) generalizes to the Fokker-Planck equation. As in Sec. 2.1, the density function for the dynamical system is defined with respect to t and $\mathbf{z}$ as $\rho(t, \mathbf{z})$. The Fokker-Planck equation is:

$$\frac{\partial \rho(\mathbf{z}, t)}{\partial t} + \nabla_\mathbf{z} \cdot (\rho(\mathbf{z}, t)\mathbf{v}(\mathbf{z})) - \nabla_\mathbf{z} \nabla_\mathbf{z}^\mathsf{T} : (\rho(\mathbf{z}, t)\mathbf{D}(\mathbf{z})) = 0, \tag{3.4}$$

where : represents a matrix double dot product $\mathbf{A} : \mathbf{B} = \mathrm{tr}(\mathbf{A}^\mathsf{T}\mathbf{B})$. The stationary density $\rho(\mathbf{z})$ is independent of t. Let $\rho(\mathbf{z}) \propto \exp(-H(\mathbf{z}))$, and write H as

$$H(\mathbf{z}) = U(\theta) + Q(\theta, \beta),$$

where β denotes general auxiliary variables. The marginal density for θ must equal the target density,

$$\exp\left(-U(\theta)\right) \propto \int \exp\left(-U(\theta) - Q(\theta, \beta)\right) \mathrm{d}\beta. \tag{3.5}$$

This condition is referred as the *marginalization condition*.

The stochastic process of θ generated by the stochastic differential equation (3.3) has the target distribution as its stationary distribution, if $\rho(\mathbf{z}) \propto \exp\left(-H(\mathbf{z})\right)$ satisfies the marginalization condition (3.5) and

$$\nabla \cdot (\rho \mathbf{v}) = \nabla \nabla^\mathsf{T} : (\rho \mathbf{D}). \tag{3.6}$$

To see this, first assume that ρ is stationary, then in the Fokker-Planck equation (3.4), $\partial \rho(\mathbf{z}, t)/\partial t = 0$. So Eq. (3.6) holds. This is the stationary Fokker-Planck equation, and is hereafter simply referred as the Fokker-Planck equation in this dissertation.

As long as the joint density $\rho(\mathbf{z})$ satisfies Eq. (3.6), it can be preserved by the dynamics. And because ρ satisfies the marginalization condition (3.5), the marginal density of θ, which is also preserved by the dynamics, equals the target density $\exp(-U(\theta))$.

3.2 Analyzing the Noise

It is essential to study the noise introduced by the stochastic gradient, before considering constructing better SDEs for sampling with less bias.

For the random variable r_k defined in Eq. (1.5), it is easy to show that

$$\mathbb{E}[r_k] = 0,$$
$$\mathbb{E}[r_k r_l] = \begin{cases} \dfrac{n-m}{m}, & l = k, \\ -\dfrac{1}{n-1}\dfrac{n-m}{m}, & l \neq k. \end{cases}$$

Write

$$\tilde{\mathbf{f}} = \mathbf{f} + \mathbf{s}, \quad \text{whence } \mathbf{s} = \sum_{k=1}^{n} r_k f_k.$$

We have

$$\begin{aligned} \boldsymbol{\Sigma} = \mathbb{E}[\mathbf{s}\mathbf{s}^\mathsf{T}] &= \sum_k \mathbb{E}[r_k^2]\mathbf{f}_k\mathbf{f}_k^\mathsf{T} + \sum_k \sum_{l\neq k} \mathbb{E}[r_k r_l]\mathbf{f}_k\mathbf{f}_l^\mathsf{T} \\ &= \frac{n(n-m)}{m(n-1)}\left(\sum_k \mathbf{f}_k\mathbf{f}_k^\mathsf{T} - \frac{1}{n}\sum_k\sum_l \mathbf{f}_k\mathbf{f}_l^\mathsf{T}\right) \\ &= \frac{n(n-m)}{m(n-1)}\left(\sum_k \mathbf{f}_k\mathbf{f}_k^\mathsf{T} - \frac{1}{n}\mathbf{f}\mathbf{f}^\mathsf{T}\right). \end{aligned}$$

This can be written as

$$\boldsymbol{\Sigma} = \frac{n(n-m)}{m(n-1)}\sum_k (\mathbf{f}_k - \mathbf{f})(\mathbf{f}_k - \mathbf{f})^\mathsf{T},$$

so it can be seen that the effect of the noise depends on the variation in $\mathbf{f}_k$.

Remark 1 *The matrix $\boldsymbol{\Sigma}$ has close relationship with the Fisher information matrix. The Fisher information matrix $\mathbf{F}$ is defined as*

$$\mathbf{F}(\theta) = \mathbb{E}_{\mathbf{x}|\theta}[\nabla_\theta l(\mathbf{x}|\theta)(\nabla_\theta l(\mathbf{x}|\theta))^\mathsf{T}],$$

where $l(\mathbf{x}|\theta) = \log(\rho(\mathbf{x}|\theta))$ and $\rho(\mathbf{x}|\theta)$ is the likelihood function. With some regularity conditions for swapping differentiation and integration,

$$\mathbb{E}_{\mathbf{x}|\theta}[\nabla_\theta l(\mathbf{x}|\theta)] = 0,$$

so

$$\mathbf{F}(\theta) = \mathrm{Var}_{\mathbf{x}|\theta}[\nabla_\theta l(\mathbf{x}|\theta)].$$

Recall that, with the prior neglected, the force $\mathbf{f}(\theta)$ *is the gradient of the negative logarithm of the likelihood function* $\rho(\mathbf{x}|\theta)$ *(see Sec. 1.1.4 and Eq. (3.1)). Therefore,* $\mathbf{\Sigma}$ *can also be written as*

$$\mathbf{\Sigma} = \frac{n(n-m)}{m}\tilde{\mathbf{F}}(\theta) \approx \frac{n(n-m)}{m}\mathbf{F}(\theta),$$

where $\tilde{\mathbf{F}}$ *is the empirical Fisher information matrix.*

Similar to the idea of stochastic gradients, the subsampled data can also be used to approximate $\mathbf{\Sigma}$. Note that

$$\mathbb{E}[\tilde{\mathbf{f}}\tilde{\mathbf{f}}^\mathsf{T}] = \frac{n(m-1)}{(n-1)m}\mathbf{f}\mathbf{f}^\mathsf{T} + \frac{n(n-m)}{m(n-1)}\sum_k \mathbf{f}_k\mathbf{f}_k^\mathsf{T},$$

and

$$\mathbb{E}[\tilde{\mathbf{f}}_k\tilde{\mathbf{f}}_k^\mathsf{T}] = \frac{n}{m}\mathbf{f}_k\mathbf{f}_k^\mathsf{T},$$

whence

$$\hat{\mathbf{\Sigma}} = \frac{n-m}{n(m-1)}\left(m\sum_k \tilde{\mathbf{f}}_k\tilde{\mathbf{f}}_k^\mathsf{T} - \tilde{\mathbf{f}}\tilde{\mathbf{f}}^\mathsf{T}\right)$$

is an unbiased estimate of $\mathbf{\Sigma}$, where $\tilde{\mathbf{f}}_k = (1 + r_k)\mathbf{f}_k$. This can be written as

$$\hat{\mathbf{\Sigma}} = \frac{n-m}{nm(m-1)}\sum_{r_k \neq -1}(m\tilde{\mathbf{f}}_k - \tilde{\mathbf{f}})(m\tilde{\mathbf{f}}_k - \tilde{\mathbf{f}})^\mathsf{T}.$$

3.2.1 The Effect of Stochastic Gradients in SDEs

When doing sampling with stochastic gradients, the full gradients in a discretized system of SDEs are replaced by the stochastic gradients. As a result, the covariance matrix $\mathbf{\Sigma}$ of the stochastic gradient introduces additional bias to the distribution of samples on top of the bias caused by discretization. The task here is to find the effect of $\mathbf{\Sigma}$ in discretized SDEs with stochastic gradients.

As a simple example, consider the Langevin dynamics written as $\mathrm{d}\mathbf{z}/\mathrm{d}t = \mathbf{v}$ with

$$\begin{aligned}\mathbf{z} &= (\theta^{\mathsf{T}}, \mathbf{p}^{\mathsf{T}})^{\mathsf{T}},\\ \mathbf{v} &= (\mathbf{p}^{\mathsf{T}}, (\mathbf{f} - (A\mathbf{I} + \mathbf{B})\mathbf{p} + (2A\mathbf{I} + 2\mathbf{B})_{1/2}\eta(t))^{\mathsf{T}})^{\mathsf{T}},\end{aligned} \tag{3.7}$$

where $\eta(t) = \mathrm{d}\mathbf{w}/\mathrm{d}t$ and $\mathbf{B}$ is an $N \times N$ matrix which is to be determined. Consider the following simple numerical integrator using splitting (Eqs. (2.7) and (2.8)):

$$\begin{aligned}\mathbf{v} &= \mathbf{v}_1 + \mathbf{v}_2,\\ \mathbf{v}_1 &= (\mathbf{p}^{\mathsf{T}}, \mathbf{0}^{\mathsf{T}})^{\mathsf{T}},\\ \mathbf{v}_2 &= (\mathbf{0}^{\mathsf{T}}, (\mathbf{f} - (A\mathbf{I} + \mathbf{B})\mathbf{p} + (2A\mathbf{I} + 2\mathbf{B})_{1/2}\eta)^{\mathsf{T}})^{\mathsf{T}}.\end{aligned}$$

The equation with $\mathbf{v}_2$

$$\mathrm{d}\mathbf{p} = \mathbf{f}\mathrm{d}t - (A\mathbf{I} + \mathbf{B})\mathbf{p}\mathrm{d}t + (2A\mathbf{I} + 2\mathbf{B})_{1/2}\mathrm{d}\mathbf{w}, \tag{3.8}$$

can be solved analytically for $\mathbf{p}$:

$$\begin{aligned}\mathbf{p}_1 &= \exp(-(A\mathbf{I} + \mathbf{B})\Delta t)\mathbf{p}_0 + (A\mathbf{I} + \mathbf{B})^{-1}(\mathbf{I} - \exp(-(A\mathbf{I} + \mathbf{B})\Delta t))\mathbf{f}_1\\ &+ (\mathbf{I} - \exp(-(2A\mathbf{I} + 2\mathbf{B})\Delta t))_{1/2}\mathbf{a}.\end{aligned} \tag{3.9}$$

When using $\tilde{\mathbf{f}}$ to replace $\mathbf{f}$, additional noise with covariance matrix $\mathbf{\Sigma}$ is added to the dynamics. To compensate for the additional noise, consider the solution of the equation

$$\mathrm{d}\mathbf{p} = \tilde{\mathbf{f}}\mathrm{d}t - (A\mathbf{I} + \mathbf{B})\mathbf{p}\mathrm{d}t + (2A\mathbf{I})_{1/2}\mathrm{d}\mathbf{w}. \tag{3.10}$$

Note that when $\mathbf{f}$ is replaced by $\tilde{\mathbf{f}}$, additional noise is added to the stochastic term in Eq. (3.10). The idea here is that $-\mathbf{B}\mathbf{p}\mathrm{d}t$ damps out the additional noise. The matrix $\mathbf{B}$ can be found by making Eq. (3.10) and Eq. (3.8) have the same solution.

The solution of Eq. (3.10) is

$$\begin{aligned}\mathbf{p}_1 &= \exp(-(A\mathbf{I} + \mathbf{B})\Delta t)\mathbf{p}_0 + (A\mathbf{I} + \mathbf{B})^{-1}(\mathbf{I} - \exp(-(A\mathbf{I} + \mathbf{B})\Delta t))\tilde{\mathbf{f}}_1\\ &\quad +(\mathbf{I} - \exp(-2A\Delta t\mathbf{I}))_{1/2}\mathbf{a}\\ &= \exp(-(A\mathbf{I} + \mathbf{B})\Delta t)\mathbf{p}_0 + (A\mathbf{I} + \mathbf{B})^{-1}(\mathbf{I} - \exp(-(A\mathbf{I} + \mathbf{B})\Delta t))\mathbf{f}_1\\ &\quad +((A\mathbf{I} + \mathbf{B})^{-2}(\mathbf{I} - \exp(-(A\mathbf{I} + \mathbf{B})\Delta t))^2\mathbf{\Sigma} + \mathbf{I} - \exp(-2A\Delta t\mathbf{I}))_{1/2}\mathbf{a}'\\ &\quad +\mathcal{O}(\Delta t^2),\end{aligned} \tag{3.11}$$

where $\mathbf{a}'$ denotes N independent Gaussian random variables, just like $\mathbf{a}$. The second equality holds because the noise in $\tilde{\mathbf{f}}$ is approximated by Gaussians having the same covariance matrix. Equate (3.9) and (3.11) and obtain:

$$\begin{aligned}&(A\mathbf{I}+\mathbf{B})^{-2}(\mathbf{I}-\exp(-(A\mathbf{I}+\mathbf{B})\Delta t))^2\boldsymbol{\Sigma}+\mathbf{I}-\exp(-2A\Delta t\mathbf{I})\\&=\mathbf{I}-\exp(-(2A\mathbf{I}+2\mathbf{B})\Delta t)+\mathcal{O}(\Delta t^2).\end{aligned}$$

An approximate solution of the equation above is

$$\mathbf{B}=\frac{\Delta t\boldsymbol{\Sigma}}{2}+\mathcal{O}(\Delta t^2). \tag{3.12}$$

For the numerical integrator used in [6], the discretized dynamics with the stochastic force $\tilde{\mathbf{f}}$ and the damping coefficient A are:

$$\begin{aligned}\theta_1&=\theta_0+\mathbf{p}_0\Delta t,\\\mathbf{p}_1&=\mathbf{p}_0+\tilde{\mathbf{f}}(\theta_1)\Delta t-A\mathbf{p}_0\Delta t+\sqrt{2A\Delta t}\mathbf{a}.\end{aligned} \tag{3.13}$$

Using the result in Eq. (3.12), the discretized dynamics (3.13) can be approximated as

$$\begin{aligned}\theta_1&=\theta_0+\mathbf{p}_0\Delta t,\\\mathbf{p}_1&=\mathbf{p}_0+\mathbf{f}(\theta_1)\Delta t-A\mathbf{p}_0\Delta t+(2A\mathbf{I}\Delta t+\Delta t^2\boldsymbol{\Sigma})_{1/2}\mathbf{a}.\end{aligned} \tag{3.14}$$

Consider the following equation of motion for momenta $\mathbf{p}$ in the Langevin dynamics:

$$\mathrm{d}\mathbf{p}=\mathbf{f}\mathrm{d}t-\mathbf{R}\mathbf{p}\mathrm{d}t+\mathbf{S}\mathrm{d}\mathbf{w}, \tag{3.15}$$

where $\mathbf{R}$ denotes the damping coefficient and $\mathbf{S}$ denotes the diffusion coefficient. The Fokker-Planck equation (3.6) applied to Eq. (3.7) requires $\mathbf{R}=\mathbf{S}^2/2$ in order to have desired stationary density $\rho(\theta,\mathbf{p})\propto\exp(-U(\theta)-\mathbf{p}^\mathsf{T}\mathbf{p}/2)$. From Eq. (3.14), however, it can be seen that the stochastic force changes the diffusion coefficient and makes it not match the damping coefficient A in the way that preserves the desired density. Therefore, the stochastic force introduces additional bias on top of the bias that is caused by the discretization error of the numerical integrator.

3.2.2 Correcting the Langevin Dynamics with Stochastic Forces

Based on Eq. (3.14), there are two ways to reduce the effect of the additional noise introduced by stochastic gradients in Langevin Dynamics. One method is to modify the diffusion coefficient by using the estimated noise $\hat{\mathbf{\Sigma}}$:

$$\begin{aligned} \theta_1 &= \theta_0 + \mathbf{p}_0 \Delta t, \\ \mathbf{p}_1 &= \mathbf{p}_0 + \tilde{\mathbf{f}}(\theta_1)\Delta t - A\mathbf{p}_0 \Delta t + (2A\mathbf{I}\Delta t - \Delta t^2 \hat{\mathbf{\Sigma}})_{1/2}\mathbf{a}. \end{aligned} \tag{3.16}$$

Note that the force here is the stochastic force. A similar modification (Eq. (3.2)) is proposed in [6], but the estimated noise $\mathbf{B}\hat{(\theta)}$ is set to 0.

The other method is to add $\hat{\mathbf{\Sigma}}$ to the damping coefficient:

$$\begin{aligned} \theta_1 &= \theta_0 + \mathbf{p}_0 \Delta t, \\ \mathbf{p}_1 &= \mathbf{p}_0 + \tilde{\mathbf{f}}(\theta_1)\Delta t - (A\mathbf{I} + \frac{\Delta t \hat{\mathbf{\Sigma}}}{2})\mathbf{p}_0 \Delta t + \sqrt{2A\Delta t}\mathbf{a}. \end{aligned} \tag{3.17}$$

The first method (Eq. (3.16)) needs the decomposition of an $N \times N$ matrix, which can be as expensive as using the full gradient when the dimension N is large. The second method (Eq. (3.17)) avoids the costly computation, because that evaluating $\hat{\mathbf{\Sigma}}\mathbf{p}_0$ only requires inner products, thanks to the special property of $\hat{\mathbf{\Sigma}}$. Therefore, the second method is more desirable in practice.

With this simple modification, the bias from the stochastic gradient can be reduced significantly, which is shown empirically later. However, the variance of the estimate of $\mathbf{\Sigma}$ is large when the size of the subset of the data is small. To this end, it is reasonable to approximate the variance with a quantity that can be estimated more accurately. For example, a scalar $\hat{\sigma}$ can be used to replace the matrix $\hat{\mathbf{\Sigma}}$:

$$\hat{\sigma} = \mathrm{tr}(\hat{\mathbf{\Sigma}})/N.$$

So $\hat{\sigma}$ condenses the information in the spectrum of $\hat{\mathbf{\Sigma}}$. Or one can make a less dramatic change by using the diagonal of $\hat{\mathbf{\Sigma}}$ instead of a single scalar.

3.3 Stochastic Gradient Nosé-Hoover Thermostat

In the last section, $\mathbf{\Sigma}$, the covariance matrix of the stochastic force, is estimated by the subsampled data. However, the large variance of the estimate for small m is undesirable. This concern leads to another idea for reducing the bias from stochastic gradients, which is to construct a new stochastic dynamical system being able to produce the desired density even if the stochastic gradient is present. The dynamics should be able to handle the additional noise without actually estimating it.

To find such dynamical system, consider the following discretized system of SDEs:

$$\begin{aligned}\theta_1 &= \theta_0 + \mathbf{p}_0 \Delta t,\\ \mathbf{p}_1 &= \mathbf{p}_0 + \tilde{\mathbf{f}}(\theta_1)\Delta t - \mathbf{\Xi}\mathbf{p}_0 \Delta t + \sqrt{2A\Delta t}\mathbf{a},\end{aligned}$$

where $\mathbf{\Xi}$ is an $N \times N$ matrix and $\mathbf{a}$ is N independent standard Gaussian random variables. The equations above are similar to Eq. (3.17), except that here $\mathbf{\Xi}$ is to be determined. Needed is a continuous system of SDEs that is a first approximation to the discretized system of SDEs, known as the *modified* equations. The desired dynamical system is to be designed based on the continuous approximation. The reason for using continuous dynamics is that the derivation is simple with the use of Fokker-Planck equation (3.6). Assume that the bias caused by stochastic gradients is much larger than the bias caused by the discretization, which is true when $n \gg m$. It is expected that the stochastic gradients and the modifications for reducing bias work in the same way for the discretized dynamics and the continuous dynamics, modulo the discretization error.

As in Sec. 3.2.1, approximate the noise in $\tilde{\mathbf{f}}$ by Gaussians having the same covariance matrix, and obtain the modified equations

$$\begin{aligned}\mathrm{d}\theta &= \mathbf{p}\mathrm{d}t,\\ \mathrm{d}\mathbf{p} &= \mathbf{f}\mathrm{d}t - \mathbf{\Xi}\mathbf{p}\mathrm{d}t + (2\mathbf{D})_{1/2}\mathrm{d}\mathbf{w},\end{aligned} \tag{3.18}$$

where $\mathbf{D}$ is also an $N \times N$ matrix. Note that in Sec. 3.2.1, the matrix $\mathbf{D}$ is actually estimated. Due to that in the discretized dynamics, $\mathbf{f}\mathrm{d}t + (2\mathbf{D})_{1/2}\mathrm{d}\mathbf{w}$ is evaluated

as $\tilde{\mathbf{f}}\Delta t + \sqrt{2A\Delta t}\mathbf{a}$, the matrix $\mathbf{D}$ in Eq. (3.18) is considered as unknown. As is discussed in Sec. 3.2.1, according to the Fokker-Planck equation (3.6), when $\mathbf{\Xi} \neq \mathbf{D}$, the dynamics do not generate desired stationary distribution $\rho(\theta, \mathbf{p}) \propto \exp(-U(\theta) - \mathbf{p}^\mathsf{T}\mathbf{p}/2)$. Therefore, the idea is to make $\mathbf{\Xi}$ a variable and find an equation of motion for $\mathbf{\Xi}$, such that the entire dynamical system guarantees the stationary distribution being the desired distribution, which should be marginalized to $\rho(\theta) \propto \exp(-U(\theta))$, no matter what the diffusion coefficient is in Eq. (3.18).

Let the equations of motion for $\mathbf{\Xi}$ be

$$\mathrm{d}\mathbf{\Xi} = \mathbf{V}_{(\mathbf{\Xi})}\mathrm{d}t, \tag{3.19}$$

which is to be determined. Let $\rho(\theta, \mathbf{p}, \mathbf{\Xi}) \propto \exp(-H(\theta, \mathbf{p}, \mathbf{\Xi}))$ be the target distribution, where H has the form

$$H(\theta, \mathbf{p}, \mathbf{\Xi}) = U(\theta) + Q(\mathbf{p}, \mathbf{\Xi}), \tag{3.20}$$

and $Q(\mathbf{p}, \mathbf{\Xi})$ is to be determined too. Clearly, the marginalization condition (3.5) is satisfied for such $H(\theta, \mathbf{p}, \mathbf{\Xi})$.

Let $R_\mathbf{z}$ denote the gradient of a function R, and $R_{\mathbf{zz}}$ denote the Hessian. For simplicity, constrain $\nabla_\mathbf{\Xi} : \mathbf{V}_{(\mathbf{\Xi})} = 0$, and assume that $\mathbf{D}$ is constant. Then the LHS and RHS of Eq. (3.6) become

$$\begin{aligned} LHS &= (\nabla \cdot \mathbf{v} - \nabla H \cdot \mathbf{v})\rho = (-\mathrm{tr}(\mathbf{\Xi}) + \mathbf{f}^\mathsf{T}\mathbf{p} - Q_\mathbf{p}^\mathsf{T}\mathbf{f} + Q_\mathbf{p}^\mathsf{T}\mathbf{\Xi}\mathbf{p} - Q_\mathbf{\Xi} : \mathbf{V}_{(\mathbf{\Xi})})\rho, \\ RHS &= \mathbf{D} : \rho_{\mathbf{pp}} = \mathbf{D} : (Q_\mathbf{p}Q_\mathbf{p}^\mathsf{T} - Q_{\mathbf{pp}})\rho. \end{aligned}$$

To cancel the $\mathbf{f}$ terms in $LHS = RHS$, set $Q_\mathbf{p} = \mathbf{p}$, then $Q(\mathbf{p}, \mathbf{\Xi}) = \mathbf{p}^\mathsf{T}\mathbf{p}/2 + S(\mathbf{\Xi})$, which leaves $S(\mathbf{\Xi})$ to be determined. Then we have

$$-\mathbf{\Xi} : \mathbf{I} + \mathbf{\Xi} : (\mathbf{p}\mathbf{p}^\mathsf{T}) - S_\mathbf{\Xi} : \mathbf{V}_{(\mathbf{\Xi})} = \mathbf{D} : (\mathbf{p}\mathbf{p}^\mathsf{T}) - \mathbf{D} : \mathbf{I}.$$

Obviously, $S_\mathbf{\Xi} : \mathbf{V}_{(\mathbf{\Xi})}$ must be a function of $\mathbf{p}\mathbf{p}^\mathsf{T}$. Since $S_\mathbf{\Xi}$ is independent of $\mathbf{p}$, it is reasonable to make $\mathbf{V}_{(\mathbf{\Xi})}$ a function of $\mathbf{p}\mathbf{p}^\mathsf{T}$. Also, $\mathbf{D}$ must appear only in $S_\mathbf{\Xi}$, since $\mathbf{V}_{(\mathbf{\Xi})}$ is expected to be independent of the unknown $\mathbf{D}$. Finally, since we let

$\nabla_{\boldsymbol{\Xi}} : \mathbf{V}_{(\boldsymbol{\Xi})} = 0$, it is reasonable to let $\mathbf{V}_{(\boldsymbol{\Xi})}$ be independent of $\boldsymbol{\Xi}$. Combining all these observations, let $\mathbf{V}_{(\boldsymbol{\Xi})}$ be a scalar multiple of of $\mathbf{p}\mathbf{p}^\mathsf{T}$ plus a constant, and $S_{\boldsymbol{\Xi}}$ is scalar multiple of $\boldsymbol{\Xi}$ plus a constant. With some algebra, the dynamics of $\boldsymbol{\Xi}$ is obtained as

$$\mathbf{V}_{(\boldsymbol{\Xi})} = (\mathbf{p}\mathbf{p}^\mathsf{T} - \mathbf{I})/\mu, \tag{3.21}$$

where μ is a free parameter, and

$$S_{\boldsymbol{\Xi}} = (\boldsymbol{\Xi} - \mathbf{D})\mu,$$

which means

$$Q(\mathbf{p}, \boldsymbol{\Xi}) = \frac{1}{2}\mathbf{p}^\mathsf{T}\mathbf{p} + \frac{1}{2}\mu(\boldsymbol{\Xi} - \mathbf{D}) : (\boldsymbol{\Xi} - \mathbf{D}).$$

Eq. (3.21) combined with Eq. (3.18) for θ and $\mathbf{p}$ give a generalized Nosé-Hoover thermostat. The stationary distribution of ξ in the deterministic Nosé-Hoover thermostat shown in Sec. 2.3.2 is $\xi \sim \mathcal{N}(0, \mu^{-1})$. With the stochastic term $(2\mathbf{D})_{1/2}\mathrm{d}\mathbf{w}$, the distribution of $\boldsymbol{\Xi}$ changes to a matrix normal distribution $\mathcal{MN}(\mathbf{D}, \mu^{-1}\mathbf{I}, \mathbf{I})$. This indicates that the varying damping coefficient $\boldsymbol{\Xi}$ adaptively adjusts itself to the matrix $\mathbf{D}$ containing the noise from the stochastic force, since the expected value of $\boldsymbol{\Xi}$ is equal to $\mathbf{D}$.

In the derivation above, $\mathbf{D}$ is assumed to be constant. The assumption that $\mathbf{D}$ is independent of θ is reasonable when the data size is large so that the variance of θ in the posterior distribution is small. When $\mathbf{D}$ depends on θ, there is still bias in the target distribution. Therefore, in general cases, the thermostat can be considered only as an approximate solution for reducing the bias. Note that the method proposed in Sec. 3.2.2 also has bias due to the variance of the estimate of $\boldsymbol{\Sigma}$, and these two sources of bias are different.

The full dynamics of $\boldsymbol{\Xi}$ requires additional $N \times N$ equations of motion, which might be too costly when N is large. In practice, further approximations needs to be considered. For example, use a scalar ξ to replace the matrix $\boldsymbol{\Xi}$, so that there is only one additional equation. The equation of ξ can be obtained by the same procedure

as that of $\boldsymbol{\Xi}$, by replacing $D\mathbf{I}$ to $\mathbf{D}$ and letting $v_{(\xi)}$ be a scalar multiple of $\mathbf{p}^\mathsf{T}\mathbf{p}$ plus a constant. So we have

$$v_{(\xi)} = (\mathbf{p}^\mathsf{T}\mathbf{p} - N)/\mu. \tag{3.22}$$

In this case, the stationary distribution of ξ is $\mathcal{N}(D, \mu^{-1})$, if $\mathbf{D} = D\mathbf{I}$. Empirically, it can be seen that this approximation is a good balance between computation cost and accuracy.

A simple numerical integrator for the dynamics of θ, $\mathbf{p}$ and ξ with stochastic gradients, named stochastic gradient Nosé-Hoover thermostat (SGNHT) [26], is described by

$$\begin{aligned}
\theta_1 &= \theta_0 + \mathbf{p}_0\Delta t, \\
\mathbf{p}_1 &= \mathbf{p}_0 + \tilde{\mathbf{f}}(\theta_1)\Delta t - \xi_0\mathbf{p}_0\Delta t + \sqrt{2A\Delta t}\mathbf{a}, \\
\xi_1 &= \xi_0 + \frac{1}{\mu}(\mathbf{p}_1^\mathsf{T}\mathbf{p}_1 - N)\Delta t.
\end{aligned} \tag{3.23}$$

Based on the analysis above, this sampler is expected to produce samples with less bias than those obtained by the ordinary Langevin dynamics with stochastic gradients. For the parameter μ, it is reasonable to have $\mu = N$, such that the rate of temperature change is independent of the dimension of the data N.

3.4 Numerical Illustrations

The experimentation begins with a toy example to demonstrate the superiority of the proposed methods for reducing the bias caused by the stochastic gradient. Various quantities are compared numerically in this example. Then the recommended method— stochastic gradient Nosé-Hoover thermostat—is run for a series of real machine learning problems. The results show the improvement in performances brought by the better sampling method at a real application level.

3.4.1 Inferring a Gaussian Distribution

Consider the Bayesian inference of a 1D Gaussian distribution. The task is to estimate both the mean and the variance of a Gaussian distribution: let n i.i.d. examples $x_i \sim \mathcal{N}(\mu, \gamma^{-1})$ be given, and let the prior be a Gaussian-gamma distribution $\mu, \gamma \sim \mathcal{N}(\mu|0, \gamma)\mathrm{Gam}(\gamma|1, 1)$, where the Gamma distribution is defined as

$$\mathrm{Gam}(\gamma|a, b) = \frac{1}{Z}\gamma^{a-1}\exp(-\gamma b).$$

The posterior is another Gaussian-Gamma distribution with the potential energy function

$$U(\mu, \gamma) = -\frac{n+1}{2}\log\gamma + \gamma + \frac{\gamma}{2}(\mu^2 + \sum_i (x_i - \mu)^2).$$

The stochastic gradient is

$$\begin{aligned}
\nabla_\mu \tilde{U} &= (n+1)\mu\gamma - \gamma \sum_{r_i \neq -1} x_i(1 + r_i), \\
\nabla_\gamma \tilde{U} &= -(n+1)/(2\gamma) + 1 + 1/2\mu^2 + \sum_{r_i \neq -1} (x_i(1 + r_i) - \mu)^2,
\end{aligned}$$

where r_i is defined in Sec. 3.1. It can be seen that the variance of the noise depends on the values of μ and γ.

The number of data in this example is 100, and the size of the random subset is 10. The damping coefficient $A = 1$, the discretization step size $\Delta t = 0.05$, and the number of samples is 10^6. Four quantities of interest are considered and the error between the true value and the estimated value (obtained from the samples) is calculated. Specifically, define e_1: the error for the expected value of the mean $\mathbb{E}(\mu)$, e_2: the error for the expected value of the standard deviation $\mathbb{E}(1/\sqrt{\gamma})$, e_3: the error for the standard deviation of the mean $\mathrm{Std}(\mu)$, and e_4: the error for the standard deviation of the standard deviation $\mathrm{Std}(1/\sqrt{\gamma})$. The mean error and the standard deviation of the error are calculated from 12 independent runs.

Six sampling methods are compared in this experiment: 1. SGHMC: replacing the full gradient by the stochastic gradient in Langevin dynamics; 2. SGLDc: Langevin dynamics with stochastic gradients and corrected damping coefficient (matrix); 3.

Table 3.1.
The mean and the standard deviation of the error between true values and the estimated values for the four quantities of interest. All numbers in the table are to be scaled by 10^{-4}.

	e_1	e_2	e_3	e_4
SGHMC	10.9 ± 7.6	229.4 ± 119.8	1934.9 ± 135.4	2069.6 ± 178.8
SGLDc	5.1 ± 3.4	11.8 ± 7.6	12.5 ± 11.3	26.5 ± 20.0
SGLDcs	9.7 ± 9.1	114.9 ± 9.0	153.8 ± 18.3	293.7 ± 18.8
SGNHTm	4.2 ± 3.0	7.0 ± 4.7	15.1 ± 4.4	6.9 ± 3.7
SGNHT	11.2 ± 5.0	112.1 ± 12.0	143.1 ± 6.6	326.2 ± 9.3
LD	9.2 ± 5.6	12.9 ± 7.9	10.5 ± 10.2	33.6 ± 17.7

SGLDcs: the scalar version of method 2; 4. SGNHTm: the matrix version of SGNHT; 5. SGNHT: the scalar version of the method; 6. LD: Langevin dynamics with full gradients, but the damping and diffusion coefficients set to the same level as of those using stochastic gradients.

The result is shown in table 3.1. From the table, it can be seen that doing a correction on the damping coefficient, either by estimating the noise or by the thermostat, greatly reduces the error. The thermostat method seems slightly better than the estimation method. Also, the matrix version methods are better than their scalar version. Fig. 3.1 shows the marginal posterior density of μ and γ. From the plots it can be seen that SGNHT recovers the true marginal density much better than SGHMC, under various parameter settings.

Although worse than the matrix versions of methods, the scalar versions are still better than the naive method, and they are the only feasible methods for real problems, which are usually of high dimension. Overall, the scalar version of SGNHT is recommended for the use in real machine learning problems.

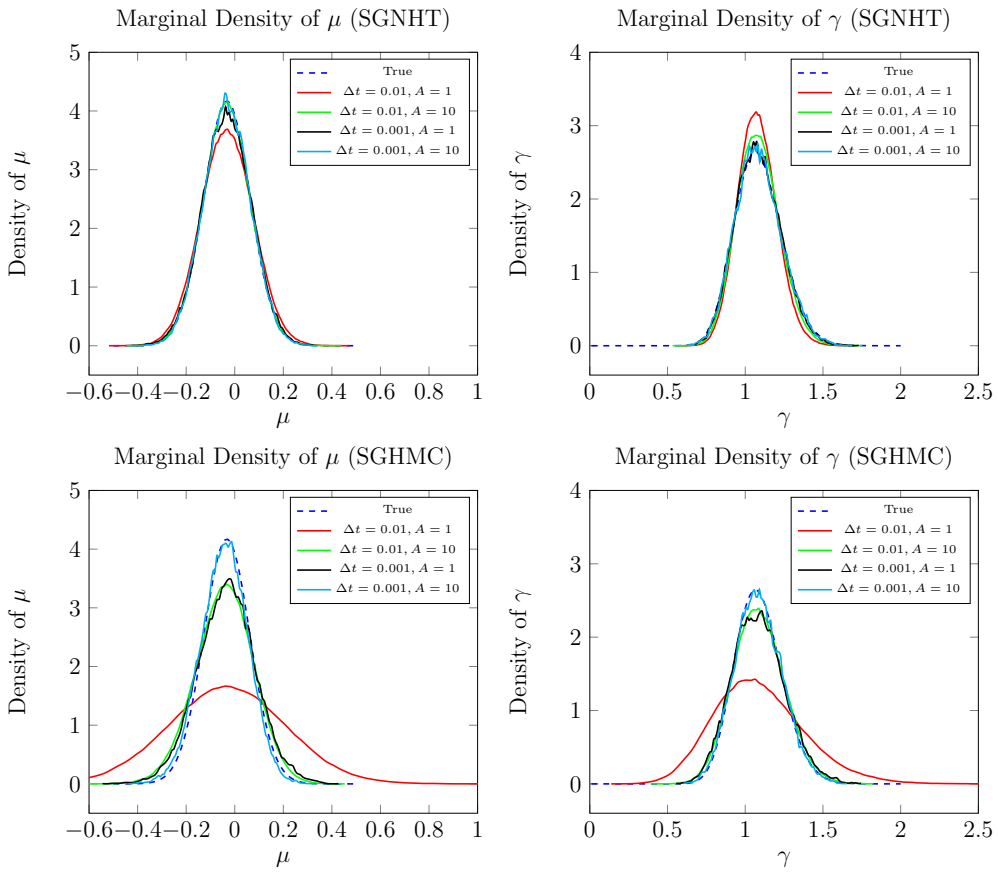


Figure 3.1. The marginal density of μ and γ recovered by SGNHT and SGHMC in the Gaussian problem.

3.4.2 Machine Learning Applications

In the following machine learning experiments, the focus is on comparing SGHMC to SGNHT. Details of the experiment settings are described below. Note that the criterion for measuring goodness of a sampler is based on the actual prediction accuracy for the Bayesian model of a problem in the experiments to be presented. The prediction accuracy for the Bayesian model generally depends on the nearness of the distribution of the samples to the true posterior distribution of the Bayesian model.

1. Bayesian Neural Network

   The first experiment is on the benchmark MNIST dataset, using the Bayesian Neural Network (BNN) as in [6]. The MNIST dataset contains 50,000 training examples, 10,000 validation examples, and 10,000 test examples. The hidden layer size of the BNN model is 100. To show SGNHT being able to handle large stochastic gradient noise, small subsets of size 20 are sampled. The total number of samples is 5×10^5 for each method. The parameter Δt^2 is chosen from $\{2, 4, 6, 8\} \times 10^{-7}$ and the product $A\Delta t$ from $\{0.001, 0.01\}$.

   Fig. 3.2 shows the testing error over various parameters. It is obvious that SGNHT outperforms SGHMC. The testing error for SGNHT is almost always smaller than SGHMC under the same parameter settings. When $\Delta t^2 = 4 \times 10^{-7}$ and $A\Delta t = 0.01$, both methods achieve the lowest testing error, and the advantage of SGNHT is the most obvious in this setting. In addition, when $A\Delta t = 0.001$, SGHMC diverges—as the green curve is way beyond the range.

2. Bayesian Matrix Factorization

   The second experiment consists of two collaborative filtering tasks: the Movielens ml-1m dataset and the Netflix dataset, using the Bayesian probabilistic matrix factorization (BPMF) model [27]. The Movielens dataset contains 6,050 users and 3,883 movies with about 1M ratings, and the Netflix dataset contains 480,046 users and 17,000 movies with about 100M ratings. The BPMF model factorizes an $N \times M$ matrix into the multiplication of an $N \times K$ matrix and a

$K \times M$ matrix. In this experiment, K is chosen to be 10. To conduct the experiments, each dataset is partitioned into training (80%) and testing (20%), and the training set is further partitioned for 5-fold cross validation. Each random subset contains 400 ratings for Movielens1M and 40,000 ratings for Netflix—0.04% of the total dataset. The total number of samples is 10^5 for each method. The parameter Δt^2 is chosen from $\{2, 4, 6, 8\} \times 10^{-7}$ and the product $A\Delta t$ is from $\{0.01, 0.1\}$ and .

Figs 3.3 and 3.4 show the root mean square error (RMSE) of the predicted rating. It can be seen that SGNHT outperforms SGHMC in all parameter settings except in the Movielens dataset with the smallest Δt^2. As Δt^2 increases, the advantage of SGNHT becomes more obvious. The divergence of SGHMC is also observed in the Netflix dataset when $\Delta t^2 \geq 6 \times 10^{-7}$ and $A\Delta t = 0.01$. In the Movielens dataset, this is even more obvious—the RMSE for SGHMC is so large that the green curve is off the chart.

3. Latent Dirichlet Allocation

   The last experiment is the ICML dataset using Latent Dirichlet Allocation (LDA) [28]. The ICML dataset contains 765 documents from the abstracts or ICML proceedings from 2007 to 2011. After simple stop-word removal, a vocabulary size of about 2K and total words of about 44K are obtained. The number of topics is 30. 80% documents are used for 5-fold cross validation and the remaining 20% for testing. Similar to [29], the semi-collapsed LDA is used. The Dirichlet prior parameter for the topic distribution for each document is set to 0.1 and the Gaussian prior for θ_{kw} is set as $\mathcal{N}(0.1, 1)$. Each random subset contains 100 documents. The total number of samples is 5×10^4 for each method.The parameter Δt^2 is chosen from $\{2, 4, 6, 8\} \times 10^{-5}$ and the product $A\Delta t$ is from $\{0.01, 0.1\}$.

   Fig. 3.5 shows the testing perplexity for various parameter settings. The testing perplexity is defined as $\mathcal{L}(\mathbf{X}_{\text{test}})^{-1/n_{\text{test}}}$, where $\mathcal{L}(\mathbf{X}_{\text{test}})$ is the likelihood on the

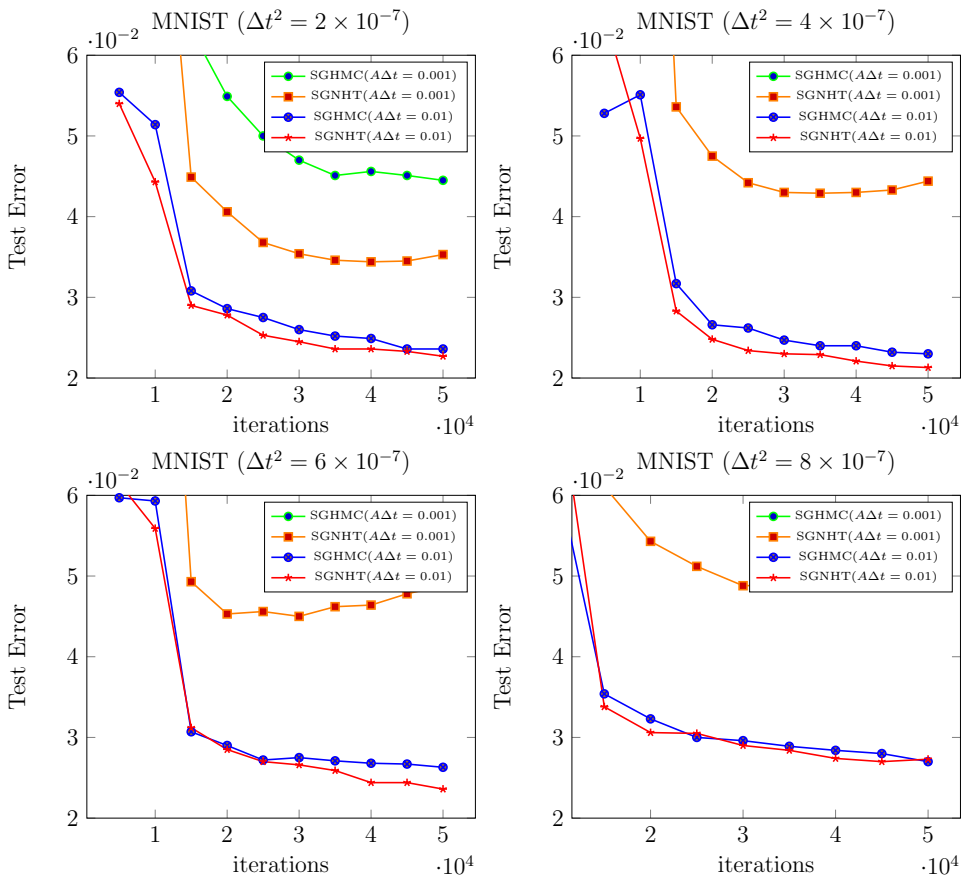


Figure 3.2. The testing error and its standard deviation for the MNIST dataset for various Δt^2 and $A\Delta t$.

testing dataset and n_{test} is the number of testing examples. It can be seen that when $\Delta t^2 = 10^{-5}$ and $A\Delta t = 0.1$, both SGNHT and SGHMC achieves the minimum perplexity, and the advantage of SGNHT is obvious. Similar to the pervious three experiments, the performance of SGHMC under some parameter settings (green curves) is too bad to be included in the chart.

Overall, SGNHT is apparently more stable than SGHMC when the discretization step Δt^2 is larger. In all four datasets, especially with the smaller $A\Delta t$, SGHMC gets worse and worse results as Δt^2 increases. Under the optimal parameter settings, SGNHT outperforms SGHMC obviously.

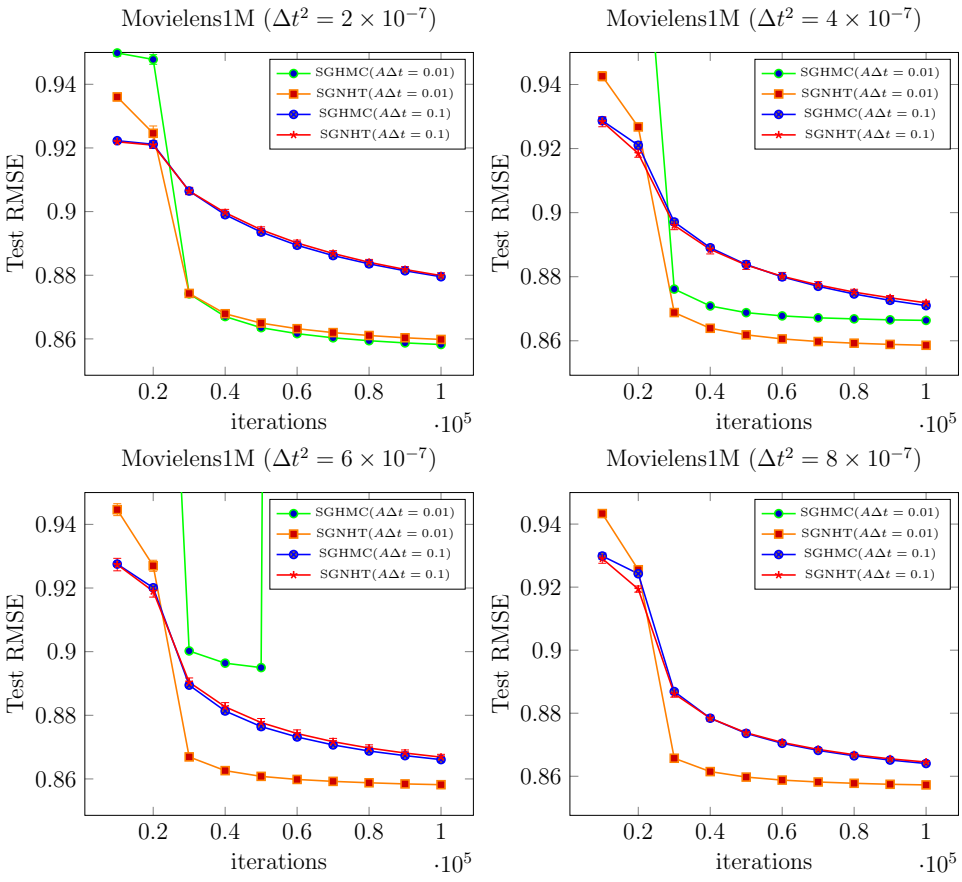


Figure 3.3. The testing RMSE and its standard deviation for the Movielens1M dataset for various Δt^2 and $A\Delta t$.

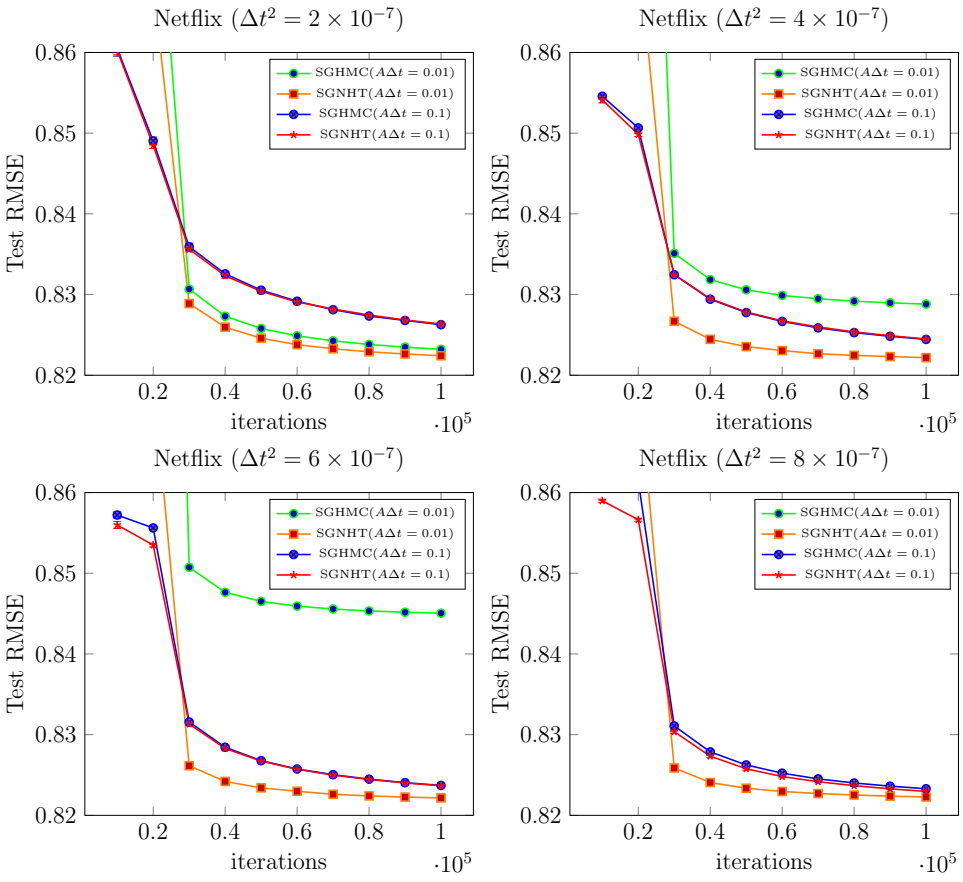


Figure 3.4. The testing RMSE and its standard deviation for the Netflix dataset for various Δt^2 and $A\Delta t$.

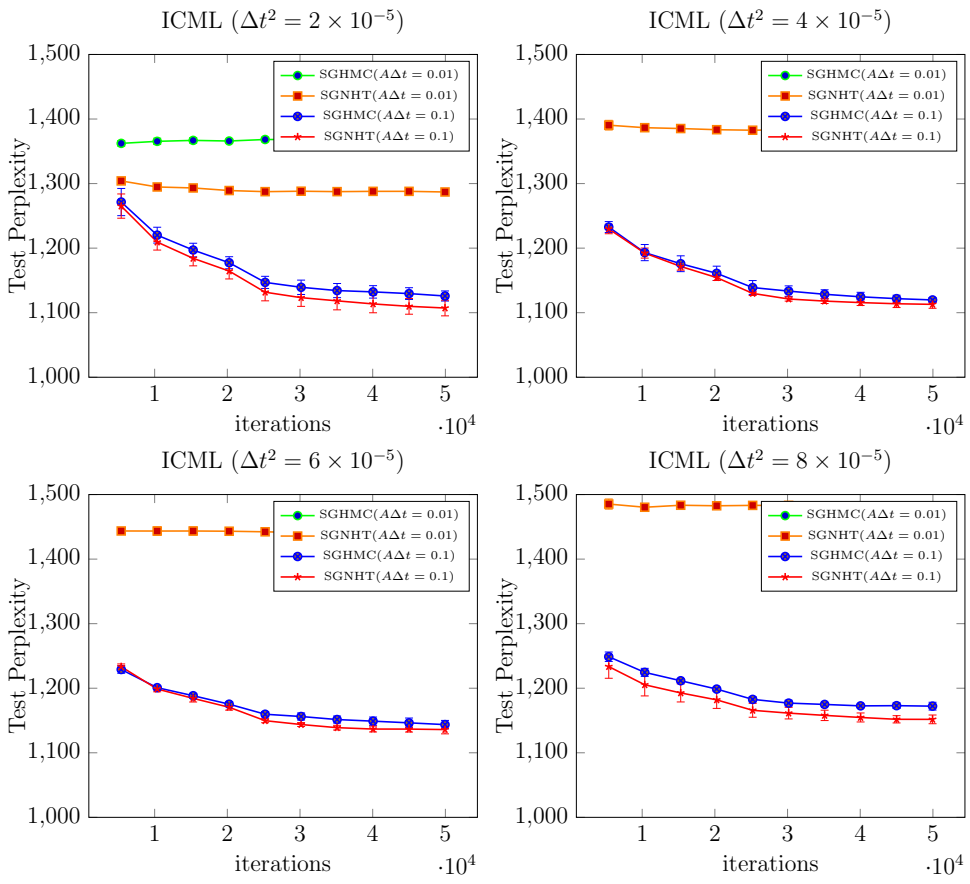


Figure 3.5. The testing perplexity and its standard deviation for the ICML dataset for various Δt^2 and $A\Delta t$.

3.5 Discussion and Conclusion

It is a challenging task to accurately sample the posterior distribution by using SDE based methods with stochastic gradients, due to the additional bias. The bias can be reduced by either estimating the noise and modifying the dynamics accordingly, or by introducing new auxiliary variables with their dynamics. Both methods are shown theoretically and empirically superior to the naive method. However, the full potential of the proposed idea cannot be fulfilled due to the costly computation involving $N \times N$ matrices, or the large variance of the estimated noise. Compromises must be made for practical use of the methods. A simple solution is to use a scalar to replace the matrix in order to reduce the computation cost and the variance. Other possible ideas include combining two methods together and using diagonal matrices instead of scalars, and applying advanced matrix computation techniques to reduce the variance of the estimate.

4 QUASI-RELIABLE ESTIMATES OF EFFECTIVE SAMPLE SIZE

Markov chain Monte Carlo methods usually require immense computational effort. So monitoring the convergence and evaluating the accuracy of MCMC methods is of prime importance. The samples generated by MCMC methods are correlated. So the effective sample size is more meaningful than the total number of samples. The effective sample size is the number of equivalently independent samples for estimating one particular quantity of interest. It can be obtained by dividing the integrated autocorrelation time τ into the total number of samples T (Eq. (1.7)).

The integrated autocorrelation time is always considered in terms of a particular quantity of interest, and it is estimated based on available samples. Then two natural questions are asked: (i) is the estimate accurate, especially when the problem is difficult such that the effective sample size is very small? (ii) is the quantity of interest representative enough for the sampler to tell if the sampling is sufficient for other quantities of interest? In fact, the number of samples needed to make accurate estimation of a quantity of interest cannot be determined solely on the estimation of the autocorrelation time of the quantity itself. Also, the autocorrelation time of one particular quantity sometimes gives false information about the efficiency of the sampler in general cases. More details about these assertions is presented later. Overall, traditional way of estimating the autocorrelation time does not give good answers to these two questions, hence is deemed unreliable for practical uses.

In this chapter, a more reliable estimate of the effective sample size is proposed, particularly for difficult problems, such as sampling multimodal distributions. In general, however, it is impossible to detect convergence without additional information. So a realistic goal is to find a quasi-reliable estimate—meaning an estimate based on apparent good coverage of state space. More concretely, it means guaranteeing the region "belonging to" the modes that have already been visited is thoroughly

sampled, and minimizing the risk of failing to visit other regions, for example, those separated by energy barriers.

Two contributions are made towards more reliable estimates of effective sample sizes:

1. An intuitive and easy to estimate quantity, $\tau_{\max}$, for measuring thoroughness of sampling is proposed. This quantity is defined to be the longest autocorrelation time over all possible functions in state space. The quantity of interest corresponding to $\tau_{\max}$ can be considered as the worst case quantity for a sampler, hence it is conservative—in a good way—for detecting the convergence.

2. An efficient algorithm for estimating this longest autocorrelation time $\tau_{\max}$ is constructed. Here efficient means that the algorithm has much less computation cost comparing to the sampling procedure. The algorithm requires only a method for estimating the integrated autocorrelation time for an arbitrary function, which is readily available and a better version can still be developed.

The idea of using the longest integrated autocorrelation time among a set of indicator functions is initially proposed in [30], and discussion about how to choose such functions automatically based on the dynamics of the propagator is presented in [31]. The method described in this dissertation is a more thorough way of approximating the function corresponding to $\tau_{\max}$. Specifically, the approximation is done by optimizing the linear combination of a set of basis functions. Note that the basis functions are not necessarily indicator functions. Also, the maximum over all linear combinations can be arbitrarily larger than the maximum over individual basis functions, if the basis functions are chosen to be highly correlated. Therefore, the proposed $\tau_{\max}$ has a better claim to the maximum.

A classical method for estimating integrated autocorrelation time of a single function is described in [7], and its implementation, a small program called `acor` is available in [32]. In this chapter, an improved method by using a better lag window, a

set of weights on combining autocorrelation functions, is also presented. Experiments show that the proposed method is slightly better than `acor`.

As a more reliable indicator of the accuracy of sampling, the longest integrated autocorrelation time is also a good measurement of the efficiency of samplers. An application of this metric is to optimize parameters of samplers to minimize $\tau_{\max}$. For example, a good choice of damping coefficient of Langevin dynamics is essential for the efficiency of the sampler. In this chapter, an initial attempt for optimizing the damping coefficient of Langevin dynamics is made. The analysis is based on a test problem where the analytical form is obtained for $\tau_{\max}$ and the optimal damping coefficient, the latter result relying on computational evidence.

The rest of this chapter is organized as follow: In Sec. 4.1, needed concepts and methods are reviewed. In Sec. 4.2, the concept of $\tau_{\max}$ is presented, and the algorithm for estimating it is described. In Sec. 4.3, a lag window that helps in getting better estimates of τ is proposed. In Sec. 4.4, theoretical and numerical analysis for the optimal parameter value for Langevin dynamics is presented. In Sec. 4.5, new methods are tested and compared with existing methods on 4 examples, covering a wide range of problems. Sec. 4.6 concludes with a discussion of the chapter.

4.1 Background

Recall from Sec. 1.1.5, the variance of the estimated mean for a quantity of interest $\mathbb{E}[u(\theta)]$ is

$$\mathrm{Var}[\widehat{u}] = \frac{1}{T}\mathrm{Var}[u(\theta)]\left(1 + 2\sum_{i=1}^{T-1}\left(1 - \frac{i}{T}\right)\frac{C(i)}{C(0)}\right).$$

As $T \to \infty$,

$$\mathrm{Var}[\widehat{u}] = \frac{1}{T}\mathrm{Var}[u(\theta)]\tau + \mathcal{O}(\frac{1}{T^2})$$

The integrated autocorrelation time τ is

$$\tau = 1 + 2\sum_{i=1}^{+\infty}\frac{C(i)}{C(0)}. \tag{4.1}$$

An estimate of covariances is provided by Ref. [33, pp. 323–324]:

$$\widehat{C}(k) = \frac{1}{T}\sum_{i=0}^{T-k-1}(u(\theta_i) - \widehat{u})(u(\theta_{i+k}) - \widehat{u}).$$

4.1.1 Forward Transfer Operator

Consider an MCMC sampler with extended space variable $\mathbf{z}$. The forward transfer operator $\mathcal{F}$ of the MCMC propagator maps a relative density $u_{i-1} = \rho_{i-1}/\rho$ for $\mathbf{z}_{i-1}$ to a relative density $u_i = \rho_i/\rho$ for $\mathbf{z}_i$:

$$u_i = \mathcal{F}u_{i-1},$$

with

$$\mathcal{F}u_{i-1}(\mathbf{z}) = \frac{1}{\rho(\mathbf{z})}\int \rho_{\mathrm{t}}(\mathbf{z}|\mathbf{z}')u_{i-1}(\mathbf{z}')\rho(\mathbf{z}')\mathrm{d}\mathbf{z}'.$$

where $\rho_{\mathrm{t}}(\mathbf{z}|\mathbf{z}')$ is the transition probability density [24]. Note that $u \equiv 1$ is an eigenfunction of $\mathcal{F}$ for eigenvalue $\lambda_1 = 1$.

Let the initial probability density be $\rho_0(\mathbf{z})$, then the error in the probability density for estimating means is,

$$\frac{1}{T}\sum_{i=0}^{T-1}\rho_i - \rho = \rho\frac{1}{T}(1-\mathcal{F}_0)^{-1}(1-\mathcal{F}_0^T)\frac{\rho_0-\rho}{\rho},$$

where operator $\mathcal{F}_0$ is $\mathcal{F}$ with the dominant eigenvalue 1 removed, i.e., $\mathcal{F}_0 u = \mathcal{F}u - \mathbb{E}[u(\theta)]$. From this, it can be seen that the error is proportional to the reciprocal of the spectral gap. In general, however, the spectral gap is not the relevant quantity for our task.

4.1.2 Error in Estimates of Integrated Autocorrelation Time

With $\hat{C}(i)$ replacing $C(i)$, an estimate of τ based on Eq. (4.1) has a standard deviation that grows with the number of terms taken; more specifically, it is approx-

imately $\sqrt{M/T}\tau$ where M is the number of terms [7, Eq. (3.19)]. Therefore, use instead

$$\widehat{\tau} \approx 1 + 2\sum_{i=1}^{T-1} w(i)\frac{\widehat{C}(i)}{\widehat{C}(0)} \tag{4.2}$$

where the weight function $w(i)$, called a *lag window*, is a decreasing function.

The program `acor` uses a lag window that is 1 for k from 0 to $M-1$ and 0 for the rest, where M is the smallest number that exceeds the estimated τ by a factor of 10. It requires the number of samples T to exceed 100τ.

4.2 The Longest Autocorrelation Time

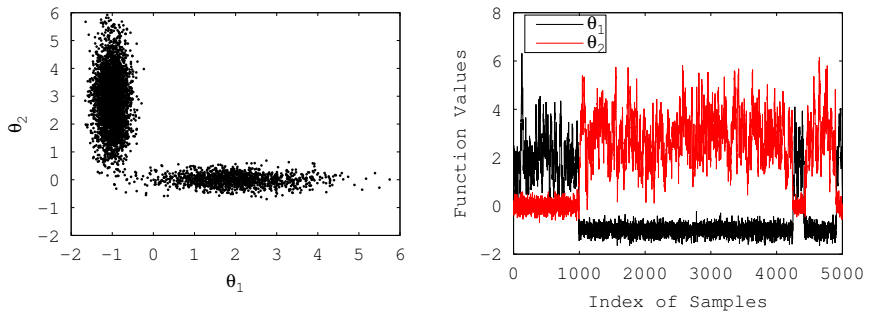
Let us begin with an example showing the danger of relying solely on estimates of the ESS for just the quantities of interest.

Consider a mixture of two Gaussians,

$$-\log\rho(\theta_1,\theta_2) = \frac{1}{2}(36(\theta_1+1)^2 + (\theta_2-3)^2) + \frac{1}{2}((\theta_1-2)^2 + 36\theta_2^2) + \text{const}, \tag{4.3}$$

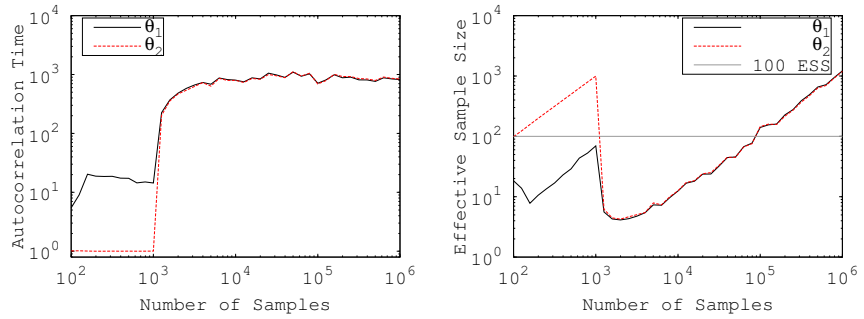
whose combined basin has an L shape with a barrier at the corner of the L. Using Brownian dynamics with the Euler-Maruyama integrator, the trajectory of a realization of sampling is shown in Fig. 4.1. It can be seen that the trajectory makes the first transition around $T = 1000$.

Fig. 4.2 shows the estimated integrated autocorrelation time and the effective sample size as a function of the sample size. In terms of thorough sampling, suppose that $ESS = 100$ is considered to be adequate for estimating quantities of interest. Just before the θ_1 curve jumps, the estimated effective sample size, $\widehat{ESS}$, is much greater than 100 indicating thorough sampling. This is misleading since after the jump $\widehat{ESS}$ drops dramatically. On the other hand, the θ_2 curve more reliably detects thorough sampling. Note that if the trajectory starts from the vertical leg of the "L", the behavior of θ_1 and θ_2 are swapped. This implies that a general quantity of interest that is independent of the starting point and can capture the "hardest" move is needed for better detecting convergence.



(a) 5000 samples (b) Trajectories of θ_1 and θ_2

Figure 4.1. Example trajectory for the L shape mixture of Gaussians distribution.



(a) Autocorrelation time (b) Effective sample size

Figure 4.2. The estimated autocorrelation time and the effective sample size vs. the number of samples in the L shape mixture of Gaussians problem.

4.2.1 Definition of $\tau_{\max}$

The autocovariance function $C(i)$ can be written in the form of a inner product. The inner product is

$$\langle v, u\rangle = \int \overline{v(\mathbf{z})} u(\mathbf{z}) \rho(\mathbf{z})\, \mathrm{d}\mathbf{z}, \tag{4.4}$$

where the bar denotes complex conjugate. Note that $\langle 1, u\rangle$ is the expectation $\mathbb{E}[u(\mathbf{z})]$ of $u(\mathbf{z})$. It can be shown by induction that the cross covariance

$$\mathbb{E}[v(\mathbf{z}_0)u(\mathbf{z}_i)] = \langle \mathcal{F}^i v, u\rangle,$$

and, in particular, $C(i) = \langle \mathcal{F}^i u, u\rangle$.

The goal is to maximize the autocorrelation time over all possible functions u. There may be permutations P of the variables $\mathbf{z}$ such that $\rho(P\mathbf{z}) = \rho(\mathbf{z})$ and $\mathcal{P}^{-1}\mathcal{F}\mathcal{P} = \mathcal{F}$, where $(\mathcal{P}u)(\mathbf{z}) = u(P\mathbf{z})$, for all interesting u. Such symmetries exist in many practical problems. For example, in machine learning, the nodes in hidden layers of a deep neural network are considered interchangeable. And in mixture models, the order of the components are also interchangeable.

For convenience, it is appropriate to consider only those functions that satisfy the symmetry condition. Define the set of function

$$W = \{u = u(\mathbf{z}) \mid \mathbb{E}[u(\mathbf{z})] = 0, u(P\mathbf{z}) = u(\mathbf{z}) \text{ for symmetries } P\},$$

and define the longest autocorrelation time as

$$\tau_{\max} = \sup_{u\in W} \left(1 + 2\sum_{i=1}^{+\infty} \frac{C(i)}{C(0)}\right).$$

By using inner product, it can also be written as

$$\tau_{\max} = \sup_{u\in W} \left(1 + 2\sum_{i=1}^{+\infty} \frac{\langle \mathcal{F}^i u, u\rangle}{\langle u, u\rangle}\right). \tag{4.5}$$

For reversible samplers, for which the propagator has a self-adjoint operator, the spectral gap is intimately related to $\tau_{\max}$. If $\tau_{\max}$ were the maximum over *all* $u(\mathbf{z})$, then

$$\tau_{\max} = (1 + \lambda_2)/(1 - \lambda_2), \tag{4.6}$$

where λ_2 is the eigenvalue of the transfer operator that is nearest to 1.

4.2.2 Discretization of State Space

Computationally, it is not possible to work in an infinite-dimensional function space, so a discretization is required. Here a linear combination $u(\theta) = \mathbf{a}^\mathsf{T}\mathbf{u}(\theta)$ of given basis functions $u_i \in W$ is considered, where $\mathbf{a}$ is to be determined. The $\tau_{\max}$ obtained by the discretized function space is expected to be close to the true value, if the basis functions are well selected.

After the discretization, the autocovariance can be written as

$$C(i) = \langle \mathcal{F}^i u, u \rangle = \mathbf{a}^\mathsf{T}\mathbf{C}_i\mathbf{a},$$

where

$$\mathbf{C}_i = \langle \mathcal{F}^i \mathbf{u}, \mathbf{u}^\mathsf{T} \rangle = \mathbb{E}[\mathbf{u}(\theta_0)\mathbf{u}(\theta_i)^\mathsf{T}], \tag{4.7}$$

and

$$\tau_{\max} \approx \sup_{\mathbf{a}} \frac{\mathbf{a}^\mathsf{T}\mathbf{K}\mathbf{a}}{\mathbf{a}^\mathsf{T}\mathbf{C}_0\mathbf{a}}, \tag{4.8}$$

where

$$\mathbf{K} = \mathbf{C}_0 + 2\sum_{i=1}^{+\infty} \mathbf{C}_i. \tag{4.9}$$

This optimization problem in Eq. (4.8) can be transformed to a generalized eigenvalue problem, by fixing the denominator and combining it with a Lagrange multiplier. After some algebra, we have

$$\frac{1}{2}(\mathbf{K} + \mathbf{K}^\mathsf{T})\mathbf{a} = \mathbf{C}_0\mathbf{a}\tau. \tag{4.10}$$

For basis functions, linear functions $u_i(\theta) = \theta_i$ are suggested as a general choice.

The matrix $\mathbf{C}_0$ is symmetric positive definite, if the components of $\mathbf{u}$ are linearly independent. For reversible samplers the cross covariance matrices $\mathbf{C}_i$, $i \geq 1$ are also symmetric [34]. A reversible MCMC sampler is defined to be one that satisfies detailed balance (Eq. (2.3)). All Metropolis-Hasting samplers are reversible. Examples of some popular reversible samplers include

1. Gibbs sampler,

2. a Brownian sampler (Sec. 1.1.3), the Euler-Maruyama discretization of Brownian dynamics coupled with a Metropolis step (known as MALA in the statistics literature [25]),

3. hybrid Monte Carlo [2] (Sec. 1.1.2).

4.2.3 A Method for Estimating $\tau_{\max}$

Begin with a guess for $\mathbf{a}$, e.g., choose $\mathbf{a}$ to be a vector of 0's except for a 1 corresponding to the function with the longest autocorrelation time. Then proceed as follow:

1. With $C(i) = \mathbf{a}^\mathsf{T}\mathbf{C}_i\mathbf{a}$, select a lag window $w(i)$, $i = 1, 2, \ldots, T-1$,

2. Set
$$\mathbf{K} = \mathbf{C}_0 + 2\sum_{i=1}^{T-1} w(i)\mathbf{C}_i. \tag{4.11}$$
Note that in practice, the lag window $w(i)$ cuts off to 0 well before $i = T-1$.

3. Choose $\mathbf{a}$ to maximize $\mathbf{a}^\mathsf{T}\mathbf{K}\mathbf{a}/(\mathbf{a}^\mathsf{T}\mathbf{C}_0\mathbf{a})$.

4. Repeat.

The number of samples T needed for an estimate of $\tau_{\max}$ depends on $\tau_{\max}$ itself; for this a formula suggested by others can be used, e.g., `acor` requires $n \geq 100\tau$, roughly.

To improve the convergence of the algorithm, $\tau_{\max}$ is prohibited from decreasing from one iteration to the next. To guarantee convergence, each $w(k)$ of the lag window is insisted to be nonincreasing from one iteration to the next.

It is important to keep the method cheap, i.e., its time complexity must not exceed the time complexity for sampling, and using FFT to estimate autocovariances is helpful for achieving this.

4.3 Better Estimates of Integrated Autocorrelation Time

In this section, a lag window for estimating τ and $\tau_{\max}$ (Eqs. (4.2) and (4.11)) is constructed.

4.3.1 An Optimal Lag Window

The autocorrelation function can be highly oscillatory, which would invalidate the method proposed here. To obtain a much more suitable function, use values based on doubling: $v_i = u(\theta_{2i}) + u(\theta_{2i+1})$. In the case of a reversible sampler, this can be shown to yield a positive decreasing convex autocorrelation function [8, Sec. 3.3]. It is straightforward to show that the autocorrelation time of u can be recovered from that of v:

$$\tau^{(u)} = \frac{1}{2}\frac{C^{(v)}(0)}{C^{(u)}(0)}\tau^{(v)},$$

where the superscripts indicate that the values belong to u or v.

4.3.2 Form of Lag Window

Begin by requiring that the lag window $w(i)$ satisfy $0 \le w(i) \le 1$ and be nonincreasing. Make the simplifying assumption that covariance estimates $\widehat{C}(i)$ satisfy

$$\widehat{C}(i) = C(i) + \sigma C(0)\eta_i,$$

for some nonnegative value σ, where the η_i are independent standard Gaussian random values. The estimated autocorrelation time is

$$\widehat{\tau} = 1 + 2\sum_{i=1}^{+\infty} w(i)\frac{c(i) + \sigma\eta_i}{1 + \sigma\eta_0},$$

where $c(i) = C(i)/C(0)$ is the normalized autocorrelation function (ACF). Assuming $\sigma \ll 1$, the error in the integrated autocorrelation time, to a first approximation, is

$$-2R + 2\sigma\sum_{i=1}^{+\infty} w(i)\eta_i - 2\sigma\eta_0\sum_{i=1}^{+\infty} w(i)c(i),$$

where

$$R = \sum_{i=1}^{+\infty} (1 - w(i))c(i).$$

The expectation of the square of the error is

$$4R^2 + 4\sigma^2 \sum_{i=1}^{+\infty} w(i)^2 + 4\sigma^2 \left(\sum_{i=1}^{+\infty} w(i)c(i) \right)^2,$$

which is minimized if

$$w(i) = \frac{c(i)}{\sigma^2} \left(R - \sigma^2 \sum_{i=1}^{+\infty} w(i)c(i) \right). \tag{4.12}$$

Note that, for small enough σ, the right-hand side is positive.

To use the formula above, the values of $c(i)$ need to be known. Therefore, consider a fitting model

$$\widehat{C}(i) = c_0 \lambda^i + \sigma c_0 \eta_i,$$

where λ, c_0, and σ are to be determined. An algorithm of finding these parameters can be found in [34]. This model is simple and well describes the tail behavior of the ACFs. For the normalized ACFs $c(i)$'s, the model assumes $c(i) = \lambda^i$, so $w(i)$ is proportional to λ^i, suggesting the choice

$$w(i) = \min\{1, \lambda^{i-M}\}, \tag{4.13}$$

where M is to be determined optimally.

Using the specified model for $\widehat{C}(i)$, the expectation of the error squared becomes

$$\frac{4c_0\lambda^2}{(1-\lambda^2)^2} \left(\mu^2 + \sigma^2(1 + \lambda - \mu)^2 \right) + \frac{4\sigma^2 \log \mu}{\log \lambda} - \frac{4\sigma^2}{1-\lambda^2},$$

where $\mu = \lambda^M$. The necessary condition for the minimum—the first derivative w.r.t. μ being 0—leads to a quadratic equation for μ, which has a positive and negative root. If the sample size T is not large enough, the positive root may exceed 1.

4.4 Optimal Damping Coefficient of Langevin Dynamics

Recall that in the equations of motion for Langevin dynamics (Eq. (1.3)), the damping coefficient A (Sec. 3.2.1) is a parameter determining the efficiency of the

sampler. The efficiency is best measured by a quantity that is independent of particular quantities of interest, since their choice varies depending on the user. Thus, the longest integrated autocorrelation time $\tau_{\max}$ is a good measurement for this purpose.

Consider a test problem—sampling 1-D Gaussian distribution $\rho(\theta) \propto \exp(-\omega^2\theta^2/2)$ by using Langevin dynamics. The equations of motion are written as

$$\begin{aligned} \mathrm{d}\theta &= p\mathrm{d}t, \\ \mathrm{d}p &= -\omega^2\theta\mathrm{d}t - Ap + \sqrt{2A}\mathrm{d}w. \end{aligned} \tag{4.14}$$

Let $\mathcal{L}_0$ be the Fokker-Planck operator (Ref. [35] and Eq. (3.4))

$$\mathcal{L}_0 u = (-p\frac{\partial}{\partial\theta} + \omega^2\theta\frac{\partial}{\partial p} + A\frac{\partial}{\partial p}p + A\frac{\partial^2}{\partial p^2})u$$

The Fokker-Planck operator is related to the forward transfer operator as

$$\mathcal{F}u = \rho^{-1}\exp(\Delta t\mathcal{L}_0)(\rho u),$$

where ρ is the stationary density.

$$\rho(\theta, p) \propto \exp(-\frac{1}{2}\omega^2\theta^2 - \frac{1}{2}p^2).$$

Therefore,

$$\mathcal{F} = \exp(\Delta t\mathcal{L}), \tag{4.15}$$

where the operator $\mathcal{L}$ is defined to be

$$\mathcal{L}u = \rho^{-1}\mathcal{L}_0(\rho u) = (-p\frac{\partial}{\partial\theta} + \omega^2\theta\frac{\partial}{\partial p} - Ap\frac{\partial}{\partial p} + A\frac{\partial^2}{\partial p^2})u.$$

The adjoint operator of $\mathcal{L}$ is

$$\mathcal{L}^\dagger u = (p\frac{\partial}{\partial\theta} - \omega^2\theta\frac{\partial}{\partial p} - Ap\frac{\partial}{\partial p} + A\frac{\partial^2}{\partial p^2})u.$$

Adjoint means $\langle\mathcal{L}v, u\rangle = \langle v, \mathcal{L}^\dagger u\rangle$.

Consider the probablists' Hermite polynomials He_i for $\omega\theta$ and p. Orthogonal basis functions of degree k can be constructed by

$$\mathrm{He}_i(\omega\theta)\mathrm{He}_j(p), \quad i + j = k.$$

The orthogonality is with respect to the inner product (Eq. (4.4)). For $k = 0, 1, 2, \ldots$, the basis functions are

$$\begin{aligned}
\mathbf{u} =& 1, \\
\mathbf{u} =& [\omega\theta, p]^{\mathsf{T}}, \\
\mathbf{u} =& [\omega^2\theta^2 - 1, \omega\theta p, p^2 - 1]^{\mathsf{T}}, \\
\mathbf{u} =& [\omega^3\theta^3 - 3\omega\theta, (\omega^2\theta^2 - 1)p, \omega\theta(p^2 - 1), p^3 - 3p]^{\mathsf{T}}, \\
\mathbf{u} =& [\omega^4\theta^4 - 6\omega^2\theta^2 + 3, (\omega^3\theta^3 - 3\omega\theta)p, (\omega^2\theta^2 - 1)(p^2 - 1), \\
& \omega\theta(p^3 - 3p), p^4 - 6p + 3]^{\mathsf{T}}, \\
& \ldots
\end{aligned}$$

The operator $\mathcal{L}$ applied to a set of basis functions $\mathbf{u}$ can be transformed to a matrix $\mathbf{M}$ applied to $\mathbf{u}$:

$$\mathcal{L}\mathbf{u} = \mathbf{M}\mathbf{u},$$

with

$$\mathbf{M} = \begin{bmatrix}
0 & k\omega & 0 & \cdots & \cdots & \cdots & 0 \\
-\omega & -A & (k-1)\omega & 0 & \cdots & \cdots & 0 \\
0 & -2\omega & -2A & (k-2)\omega & 0 & \cdots & 0 \\
0 & 0 & \cdots & & & \cdots & 0 \\
\cdots & & & & -(k-1)\omega & -(k-1)A & \omega \\
0 & \cdots & & \cdots & 0 & -k\omega & -kA
\end{bmatrix},$$

where k is the number of basis functions. Similarly, the adjoint operator $\mathcal{L}^\dagger$ has a corresponding matrix $\mathbf{M}^\dagger$:

$$\mathbf{M}^\dagger = \begin{bmatrix}
0 & -k\omega & 0 & \cdots & \cdots & \cdots & 0 \\
\omega & -A & -(k-1)\omega & 0 & \cdots & \cdots & 0 \\
0 & 2\omega & -2A & -(k-2)\omega & 0 & \cdots & 0 \\
0 & 0 & \cdots & & & \cdots & 0 \\
\cdots & & & \cdots & (k-1)\omega & -(k-1)A & -\omega \\
0 & \cdots & & \cdots & 0 & k\omega & -kA
\end{bmatrix}.$$

Recall the relation between $\mathcal{L}$ and $\mathcal{F}$ (Eq. (4.15)). Combining Eqs. (4.7), (4.9) and (4.10), $\tau_{\max}$ for a given k is the maximum eigenvalue of the matrix

$$\mathbf{I} + \sum_{i=1}^{+\infty} \exp(i\mathbf{M}\Delta t) + \sum_{i=1}^{+\infty} \exp(i\mathbf{M}^\dagger \Delta t) = \coth(-\frac{1}{2}\mathbf{M}\Delta t) + \coth(-\frac{1}{2}\mathbf{M}^\dagger \Delta t).$$

Note that the maximum eigenvalue of this matrix is a function of A, ω and Δt. Actually the number of parameters can be reduced to 2, by extracting ω out of the matrices $\mathbf{M}$ and $\mathbf{M}^\dagger$. Then $\tau_{\max}\omega\Delta t$ can be made a function of A/ω. The reason to make A/ω one of the parameters is because that omega is intrinsic to the problem and Δt is just a method parameter. The longest autocorrelation time $\tau_{\max}$ can be found numerically from the matrix with different sets of basis functions. Assuming the maximum of $\tau_{\max}$ is attained for $k = 1$ and $k = 2$, which is supported by Fig. 4.3. The optimal A is at the intersection of the curve for $k = 1$ and the curve for $k = 2$. An analytical form for the optimal A is $\sqrt{6}\omega/2$ as $\Delta t \to 0$, which is derived in [16]. From Fig. 4.3, it can be seen that the numerical value is consistent with the theoretical value.

4.5 Numerical Experiments

4.5.1 A Gaussian Distribution

To confirm the correctness of the theory, consider a simple test problem for sampling where the target distribution is the standard Gaussian. The sampler is Brownian dynamics with Euler-Maruyama method. The discretization time step size is chosen to be $\Delta t = 0.02$. This step size is much smaller than needed in practice. It is chosen in this way to make the discretization error negligible, thereby permitting the use of analytical results derived for exact Brownian dynamics. The Markov chain is obtained by subsampling the original trajectory at intervals of 0.1 (every 5th point). This gives the true eigenvalues of the transfer operator of the Brownian dynamics [35] as $1, \exp(-0.1), \exp(-0.2), \exp(-0.3), \ldots$. The corresponding eigenfunctions are the

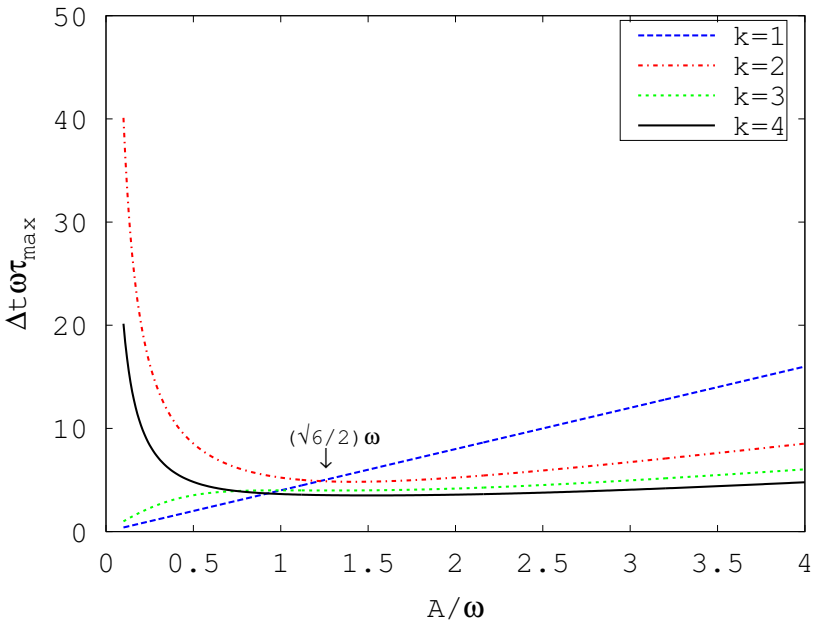


Figure 4.3. $\Delta t\omega\tau_{\max}^{(k)}$ vs. A/ω for $k = 1, 2, 3, 4$, $\Delta t = 10^{-6}$ and $\omega = 1$.

Table 4.1.
The weights of the linear combination of the three basis functions with a_1 normalized to 1.

	$n = 10^3$	$n = 10^4$	$n = 10^5$	$n = 10^6$
a_1	-0.036	-0.008	0.003	0.010
a_2	0.972	1.024	1.008	0.994
a_3	1.000	1.000	1.000	1.000

Hermite polynomials defined by the recurrence: $\mathrm{H}_i(\theta) = 2q\mathrm{H}_{i-1}(\theta) - \mathrm{H}'_{i-1}(\theta)$ with $\mathrm{H}_0(\theta) = 1$.

Consider three basis functions

$$\mathrm{H}_3 + \mathrm{H}_2 + \mathrm{H}_1, \quad \mathrm{H}_3 - \mathrm{H}_2 + \mathrm{H}_1, \quad -\mathrm{H}_3 + \mathrm{H}_2 + \mathrm{H}_1.$$

The goal of this test is to recover the theoretical value of $\tau_{\max}$ and the corresponding maximizing function $\mathrm{H}_1(\theta)$ by using the linear combination of given functions. Since Brownian dynamics is a symmetric sampler, the theoretical value of $\tau_{\max}$ can be obtained by using Eq. (4.6):

$$\tau_{\max} = (1 + \exp(-0.1))/(1 - \exp(-0.1)) \approx 20.0.$$

Twelve independent runs are performed to evaluate the reliability of the method.

Fig. 4.4 shows the estimated $\tau_{\max}$ for all 12 runs for increasing sample size. It can be seen that the estimated value converges to the true value. And the variance is comparatively small when $N > 3 \times 10^3$. Table 4.1 shows the coefficients of the linear combinations of given basis functions. It can be seen that the theoretical maximizing function $\mathrm{H}_1(\theta)$ is successfully recovered.

The proposed lag window and the lag window of `acor` are also compared. Fig. 4.5 shows the mean and variance for 12 runs of the estimates of τ for H_1 and $\mathrm{H}_3 + \mathrm{H}_2 + \mathrm{H}_1$ as well as $\tau_{\max}$. It can be seen that when T is small, both methods underestimate τ, but the proposed method gives larger estimates. When T is large, both methods

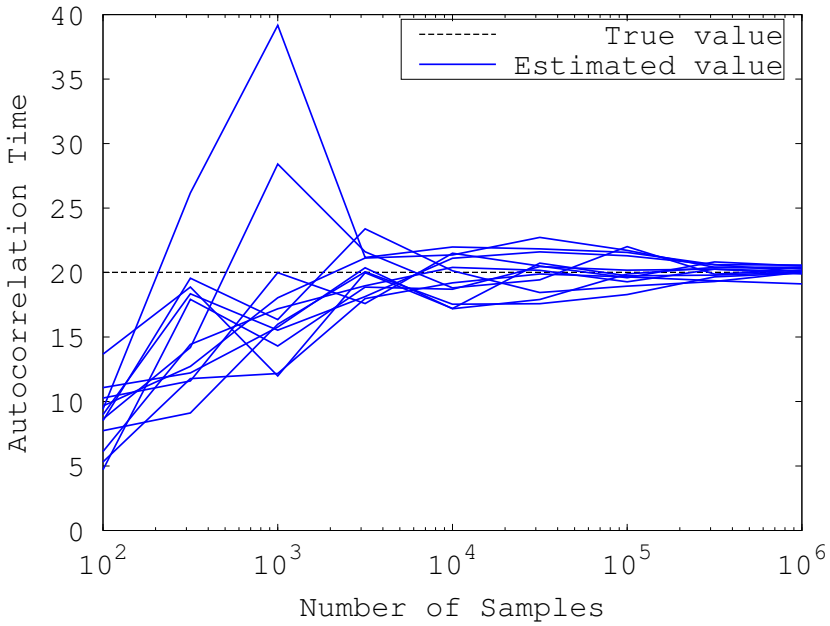
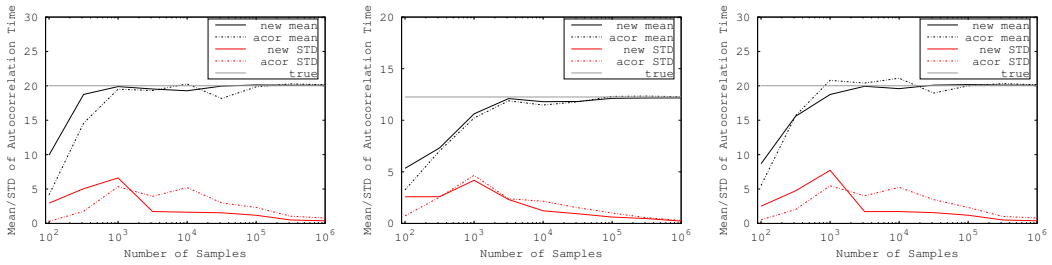


Figure 4.4. The estimated $\tau_{\max}$ in the 1-D Gaussian problem.



(a) H_1 (b) $H_3 + H_2 + H_1$ (c) $\tau_{\max}$

Figure 4.5. Estimated τ and $\tau_{\max}$ by the proposed lag window and `acor`'s lag window for the 1-D standard Gaussian.

converge to the true value, and the proposed method has a smaller variance. Overall, the proposed lag window is more reliable than is that of `acor`.

4.5.2 An L-Shaped Mixture of Two Gaussian Distributions

Consider again the mixture of two Gaussians with target density given by Eq. (4.3) sampled with discretized (Euler-Maruyama) Brownian dynamics with time step size $\Delta t = 0.02$ and subsampled at time interval 0.1 (every 5 points). The basis functions are θ_1 and θ_2. Theoretically, θ_1 should have a larger estimated autocorrelation time than θ_2 for small T while the chain remains in the horizontal leg of the "L", since the frequency of the motion is smaller in the direction of θ_1 than it is in the direction of θ_2. By the same reasoning, the weight for θ_1 in the linear combination forming the maximizing function should be greater than the weight for θ_2, when T is small. Due to symmetry, on the other hand, both autocorrelation times and the magnitudes of the weight for θ_1 and θ_2 should be equal when T is large.

Fig. 4.2 shows $\widehat{ESS}$ and $\widehat{\tau}$ for θ_1 and θ_2. Table 4.2 shows the weights of linear combinations for different sample sizes. The weighting obtained meets expectations, ultimately giving equal weight to θ_1 and θ_2. Computation of $\tau_{\max}$ (not shown due to

Table 4.2.
The weights of the linear combination of the θ_1 and θ_2 with a_1 normalized to 1.

	$n = 10^3$	$n = 10^4$	$n = 10^5$	$n = 10^6$
a_1	1.000	1.000	1.000	1.000
a_2	0.089	0.085	−1.207	−1.051

overlapping with the curve for θ_1) confirms that its estimate equals the greater the estimates of the τ values for the two basis functions.

4.5.3 A One-node Neural Network

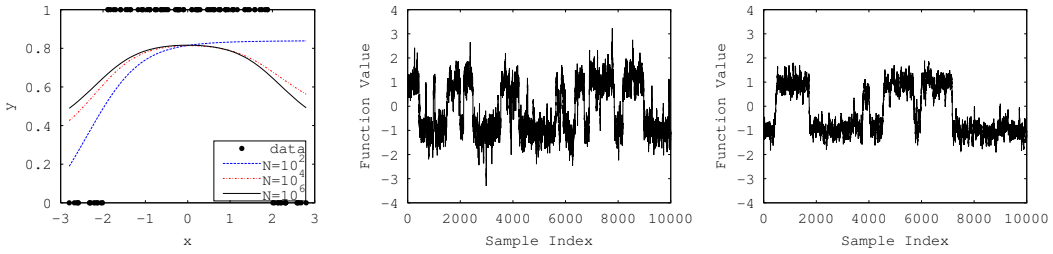
Consider the simplest Bayesian neural network regression model [36], having a single node in the hidden layer:

$$y_i \approx u(\theta; x_i) = \theta_3 \tanh(\theta_1 x_i + \theta_2) + \theta_4,$$

where (x_i, y_i) represents a data example. Suppose a total of 100 data examples are given as shown by the large dots of Fig. 4.6(a). The posterior distribution is

$$-\log \rho(\theta) = \frac{1}{2}\beta \sum_{i=1}^{100} (y_i - u(\theta; x_i))^2 + \frac{1}{2}\alpha \|\theta\|^2 + \text{const},$$

where $\beta = 2.5$ and $\alpha = 0.8$. There are three modes. One corresponds to a good fit at $x = -2$, one to a good fit at $x = 2$, and one to a best fit as a constant function. Suppose the sampler is discretized (Euler-Maruyama) Brownian dynamics with $\Delta t = 0.01$, and let the Markov chain be obtained by subsampling the original trajectory at time interval 0.1 (every 10 points). Use as basis functions for finding the maximizing function by using linear combination are the parameters of the model θ_1, θ_2, θ_3 and θ_4. In practice, $\mathbb{E}_\theta[u(\theta; \mathbf{x})]$ for any $\mathbf{x}$ is a possible quantity of interest. In this experiment, only the prediction for the first data example is computed.



(a) The prediction (b) Trajectory of θ_1 (c) Trajectory of $u_{\max}$

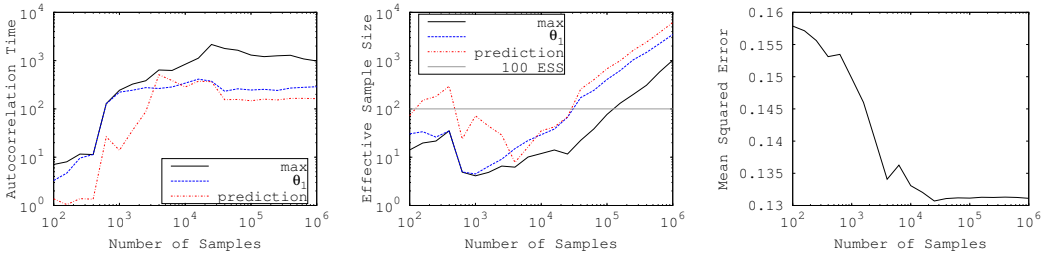
Figure 4.6. The resulting prediction and the trajectories of function values in the one-node neural network problem.

Fig. 4.6(a) shows the overall prediction $\bar{u}(x)$ resulting from the regression model. It can be seen that when the sampling is insufficient, the prediction is not the best possible— the asymmetry implies only one mode is explored, just like the behavior of the maximum a posteriori (MAP) estimate. When the sampling is sufficient, on the other hand, the prediction is much closer to being the best possible. This example demonstrates the advantage of sampling over MAP, and shows the necessity of detecting convergence to validate a sampling result.

Fig. 4.6(b) and (c) show the trajectories of function values. It can be seen that the maximizing function makes many fewer transitions than θ_1. This is because the maximizing function tends to make the "hardest" moves, hence has longest autocorrelation time.

Fig. 4.7(a) and (b) show $\widehat{ESS}$ and $\widehat{\tau}$ for the maximizing function and two others. It can be seen that $\widehat{ESS}$ of the prediction $\bar{u}(x_1)$ is misleading for convergence detection, since it indicates sufficient sampling when $n \approx 100$. It can also be seen that the maximum $\widehat{\tau}$ is significantly larger than the longest $\widehat{\tau}$ of the other two functions.

Fig. 4.7(c) shows the mean squared error of the regression converging when $N > 2.0 \times 10^4$. This coincides with the convergence of $\hat{\tau}_{\max}$. On the other hand, the $\widehat{\tau}$ for the prediction $\bar{u}(x_1)$ might suggest stopping far too early and the $\widehat{\tau}$ for θ_1 might also be too optimistic.



(a) τ (b) ESS (c) MSE

Figure 4.7. Autocorrelation times, effective sample sizes, and the mean squared error for the one-node neural network model.

4.5.4 Logistic Regression

Consider a Bayesian logistic regression model [10] discussed in Sec. 1.1. The logistic function maps a linear combination of features $\mathbf{x}$ to a real value in $(0, 1)$:

$$\sigma(\theta; \mathbf{x}) = 1/(1 + \exp(-\theta^{\mathsf{T}}\mathbf{x})).$$

The posterior distribution is

$$-\log \rho(\theta) \quad = \quad \beta \sum_{i=1}^{n} (y_i \log(\sigma_i) + (1 - y_i) \log(1 - \sigma_i)) + \frac{1}{2}\alpha \|\mathbf{q}\|^2 + \text{const},$$

where y_i is the class label of data example i, $\beta = 1.0$ and $\alpha = 0.1$. When predicting the label of a given data example $\mathbf{x}$, use the average value

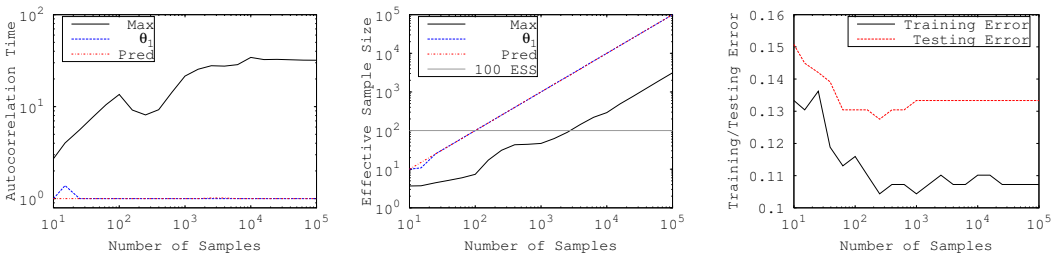
$$\bar{\sigma}(\mathbf{x}) = \frac{1}{T} \sum_{i=0}^{T-1} \sigma(\theta_i; \mathbf{x}),$$

over all samples. For data, use the Australian Credit Approval dataset from the UCI machine learning data repository [37]. It contains 690 data examples and provides 15 features for each data example. Half of the examples in the dataset are extracted to form the training set, and the remainder are from the testing set. The sampler is discretized (Euler-Maruyama) Brownian dynamics with $\Delta t = 0.05$. This step size best balances discretization error and sampling error. As for the one-node neural network problem, the model parameters θ_i are used as functions for estimating $\tau_{\max}$ and the prediction for the first data point in the training set as a quantity of interest.

Fig. 4.8 shows $\widehat{\tau}$, $\widehat{ESS}$, and the training/testing error in the logistic regression problem. The training error is defined as

$$err_{\text{train}} = \frac{1}{n} \sum_{i=1}^{n} 1_{t(\mathbf{x}_i) \neq y_i},$$

where the label t of a data example $\mathbf{x}$ is predicted to be 1 if $\bar{\sigma}(\mathbf{x}) \geq 0.5$ and 0 if $\bar{\sigma}(\mathbf{x}) < 0.5$. Note that t itself is not a quantity of interest, but depends on the quantity of interest $\mathbb{E}_\theta[\sigma(\theta; \mathbf{x})]$. The testing error is defined similarly except the data examples are from the testing set.



(a) τ (b) ESS (c) Training error

Figure 4.8. The autocorrelation times, the effective sample sizes, and the training/testing error in the logistic regression problem.

In practice, the convergence may be monitored by observing the convergence of the training error. This approach, however, can be very expensive for large data sets. This is because that the prediction function $\sigma(\theta; \mathbf{x})$ has to be evaluated for every data example $\mathbf{x}_i$'s, so that the computation cost is even higher than the sampling procedure if stochastic gradients are used. Instead, estimates of quantities, such as the prediction function of a single data example or a single weight, can be used for detecting convergence. As is shown in the figure, however, it is risky to observe only one such quantity. Both the training and testing error do not converge until the number of samples exceeds 1000, but both the prediction and the weight θ_1 have an almost constant autocorrelation time equal to 1 (implying independent samples). This is obviously misleading. On the other hand, the maximizing function does not show convergence until $N > 1000$. This demonstrates that the maximizing function is a more reliable indicator of convergence.

4.6 Discussion and Conclusion

For MCMC methods, there is no doubt of the importance of estimating the integrated autocorrelation times τ for quantities of interest that are estimated from the samples. However, estimates of τ by previous methods are not reliable. In this chap-

ter, a method that uses the results of the sampling to obtain quasi-reliable estimates is proposed. The idea is to find the function with the longest τ among all possible linear combination of a set of basis functions. The effectiveness of the method depends on how well the space spanned by the basis functions approximates the space of all functions of the state space. To select good basis functions, intuition and prior knowledge of the problem is helpful. If such information is not available, low-degree polynomials might be reasonable choices.

5 FURTHER WORK ON SAMPLING METHODS

This chapter is organized as follow: Sec. 5.1 presents the idea of changing variables, with two examples showing the utility of this idea; Sec. 5.2 describes the two-stage Takahashi-Imada method, and shows results of experiments; Sec. 5.3 concludes this chapter with a discussion.

5.1 Change of Variables

The methods proposed in Chap. 2 focus on designing dynamical systems to overcome the obstacles for efficient sampling caused by the undesirable properties of the potential energy function. For example, the variable mass method presented in Sec. 2.3.4 changes the Hamiltonian dynamics to the variable mass Hamiltonian system in order to allow the particle moving faster in the region of the energy barrier. In fact, rather than designing the dynamical system of HMC-like samplers, the difficulty of sampling could be reduced by another way—changing the potential energy function itself.

This is achieved by the idea of change of variables. The aim to change the variables θ is to change the potential energy function that governs the dynamics. In the transformed space, the undesirable properties of the potential energy function can be removed, by carefully designing the formula for the new variable. The original idea is from [9]. Such method, however, is very complicated, and extensive problem-specific knowledge is needed.

The goal of this work is to make a general framework for changing variables, under which simpler methods can be found to change the potential energy function to favor better sampling. Moreover, less problem-specific knowledge is required, such that the idea is more generally applicable. Note that the framework of designing dynamical

systems presented in Chap. 2 is independent of any prior transformation. Therefore, the method discussed in this chapter can be combined with methods in Chap. 2 to make even more powerful sampling methods.

There are two examples to be shown for the purpose of demonstration. The first one is a simple test problem—a double-well potential in 1-D. The task is to remove the barrier in the transformed space by applying a simple change of variable. The second one is a problem from machine learning—the relevance vector machine (RVM) [38]. The potential energy function of RVM has a smooth region combined with a spike region. The spike region requires a very small time step size for HMC to have a reasonable acceptance probability, while the smooth region cannot be sampled efficiently with the same step size. The task for this problem is to remove the spike region, again by a simple change of variables, such that the ordinary HMC is applicable.

5.1.1 The Method

Consider the change of variables

$$\theta = \mathbf{g}(\xi). \tag{5.1}$$

where $\mathbf{g}$ is a one-on-one mapping from ξ to θ. Assume the probability density of θ is $\rho_\theta(\theta) = \exp(-U(\theta))/Z$, and the probability density of ξ is $\rho_\xi(\xi) = \exp(-U'(\xi))/Z$. To find the potential energy function of ξ, consider an arbitrary subset $\mathcal{A}$ of the phase space,

$$\begin{aligned}
&\int_{\mathcal{A}} \frac{1}{Z} \exp(-U'(\xi))\mathrm{d}\xi = \mathbb{P}(\xi \in \mathcal{A}) = \mathbb{P}(\mathbf{g}^{-1}(\theta) \in \mathcal{A}) \\
&= \int_{\mathbf{g}(\mathcal{A})} \frac{1}{Z} \exp(-U(\theta))\mathrm{d}\theta = \int_{\mathcal{A}} \frac{1}{Z} \exp(-U(\mathbf{g}(\xi))|\det(\partial_\xi \mathbf{g}(\xi))|\mathrm{d}\xi \\
&= \int_{\mathcal{A}} \frac{1}{Z} \exp(-U(\mathbf{g}(\xi)) + \log(|\det(\partial_\xi \mathbf{g}(\xi))|))\mathrm{d}\xi
\end{aligned}$$

Therefore

$$U'(\xi) = U(\mathbf{g}(\xi)) - \log(|\det(\partial_\xi \mathbf{g}(\xi))|). \tag{5.2}$$

The original HMC algorithm is modified to the algorithm of HMC with change of variables (HMCcv) by changing the propagator from Ψ to $\tilde{\Psi}$, where

$$\tilde{\Psi} = \mathbf{g} \circ \Psi' \circ \mathbf{g}^{-1}.$$

The propagator Ψ' is the discretized flow of the equations of motion

$$\mathrm{d}\xi = \mathbf{p}\mathrm{d}t,$$
$$\mathrm{d}\mathbf{p} = \mathbf{f}'\mathrm{d}t,$$

where

$$\mathbf{f}' = (\partial_\xi \mathbf{g})^\mathsf{T}\mathbf{f}(\mathbf{g}) - \nabla_\xi \log(|\det(\partial_\xi \mathbf{g})|). \tag{5.3}$$

Note that in the actual implementation, $\mathbf{g}$ is only needed at the end of each MC step to obtain samples in the original space, and $\mathbf{g}^{-1}$ is only needed once at the beginning to transform the initial sample θ_0 to ξ_0.

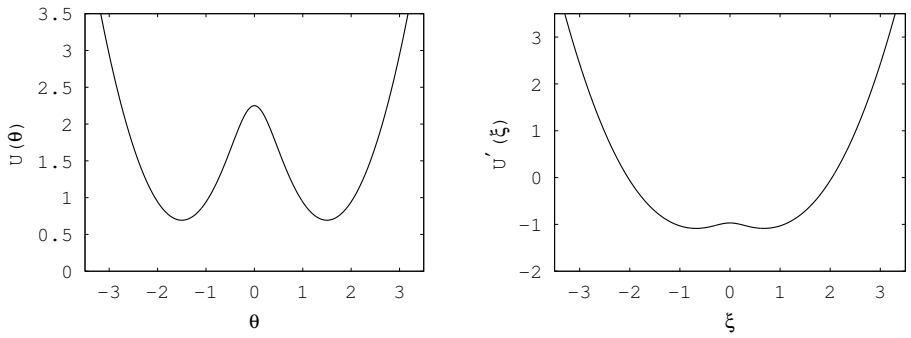
5.1.2 Lowering the Energy Barrier

To use the idea of changing variables to lower the energy barrier, a function $\mathbf{g}(\xi)$ needs to be found, such that in the space of ξ, the energy barrier is lowered and, at the same time, the function itself is kept simple. REPSWA [9] uses Legendre polynomials to approximate the potential function in the barrier region. This approximation is complicated and requires extensive expert knowledge about the problem. The goal in this section is to find a simple function that achieves the same.

Consider a 1-D double-well potential problem. To make the barrier flat, find a monotonic function $\theta = g(\xi)$ such that the second term in Eq. (5.2), $\log(\mathrm{d}g(\xi)/\mathrm{d}\xi)$, is a "bell-shaped" function located in the barrier region. In this way, by subtracting this term, the barrier in the potential energy function can be removed or lowered.

Letting $\log(\mathrm{d}g(\xi)/\mathrm{d}\xi) = l(g)$, we have

$$\frac{\mathrm{d}g}{\mathrm{d}\xi} = \exp(l(g)). \tag{5.4}$$



(a) (b)

Figure 5.1. The graphs of $U(\theta)$ (a) and $U'(\xi)$ (b) in the 1-D double-well potential problem.

Integrate Eq. (5.4) to get

$$\xi = \int_0^{g^{-1}(\xi)} \exp(-l(\theta))\mathrm{d}\theta,$$

and

$$g(\theta) = \int_0^{\theta} \exp(-l(\theta'))\mathrm{d}\theta'.$$

Here $l(\theta')$ is to be determined and must satisfy two conditions: (i) in order to have a closed form for ξ, it must be integrable analytically; (ii) it must be "bell-shaped". To this end, try

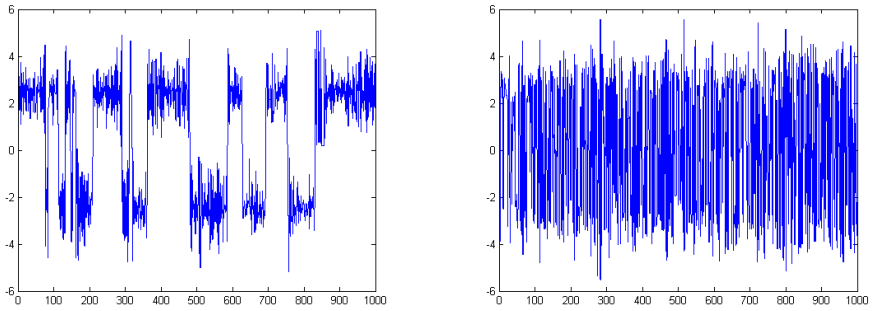
$$l(\theta') = -\log(1 - \alpha \mathrm{sech}^2(\frac{\theta'}{\sigma})),$$

where $0 < \alpha < 1$ and $\sigma > 0$ are parameters to control the shape of the function in order to fit the width and the height of the barrier. After some algebra, a formula for ξ is found to be

$$\xi = \theta - \alpha\sigma \tanh(\frac{\theta}{\sigma}). \tag{5.5}$$

Fig. 5.1 shows the graphs of $U(\theta)$ and $U'(\xi)$. It can be seen that the energy barrier is greatly reduced in the transformed variable.

The method is tested in a 1-D double-well potential problem arising from a mixture of Gaussians located at ± 1.5 with standard deviations 1. The integrator is the leapfrog



(a) (b)

Figure 5.2. The trajectory of the samples for HMC (left) and HMC with change of variables (right). The x-axis is the number of samples and the y-axis is the position.

method for Hamiltonian dynamics. Table 5.1 shows the result of the experiment. In the table, P is the acceptance probability, ε the probability of crossing the energy barrier in a given step. The parameters $\alpha = 0.96$ and $\sigma = 3$. The integration step sizes for HMC and HMCcv are $\pi/2$ and $\pi/3$, respectively. The integration duration is $\tau = 1$. The number of samples is 10^3.

Table 5.1.
The acceptance probability and the crossing probability in a given step for the method of changing variables in the 1-D double-well potential problem. P is the acceptance probability, ε is the crossing probability, Δt is the step size, HMCcv is HMC with change of variables

	P	ε	Δt
HMC	0.82	2.24%	$\pi/2$
$HMCcv$	0.80	47.71%	$\pi/3$

Fig. 5.2 shows the trajectories of θ. It can be seen that the change of variables helps a lot in crossing the energy barrier—about 20 times more events.

5.1.3 Smoothing the Potential Energy Function

In some specific models in machine learning, such as the Relevance Vector Machine (RVM) [38], the potential energy function has some spikes resulting in extremely high frequency motions. Very small step sizes are required for HMC to work in those particular regions, but in other regions, such step sizes are too small to explore efficiently. A change of variables can solve this problem by removing the spikes from the potential energy function so that uniformly larger step sizes can be applied.

The potential energy function of an RVM can be written as

$$U(\theta) = \theta^{\mathsf{T}}\mathbf{A}\theta - \mathbf{b}^{\mathsf{T}}\theta + \sum_{i=1}^{N} \frac{1}{2}\log(\alpha + \theta_i^2), \tag{5.6}$$

where $\mathbf{A}$ and $\mathbf{b}$ are obtained from the data, and α is a very small positive value close to zero. This function consists of a quadratic term and a logarithm term. The step size for HMC to sample efficiently in most of the regions in the state space fails for sampling the region corresponding to the logarithm term.

To remove the logarithm term, equate the logarithm term to the second term in U' (Eq. (5.2)):

$$\log(|\det(\partial_\xi \mathbf{g}(\xi))| = \sum_{i=1}^{N} \frac{1}{2}\log(\alpha + g_i^2).$$

Note that if each g_i is chosen to be a function of θ_i alone, we have

$$\log|\frac{\mathrm{d}g_i}{\mathrm{d}\xi_i}| = \log\sqrt{\alpha + g_i^2}.$$

After some algebra, the equation above gives

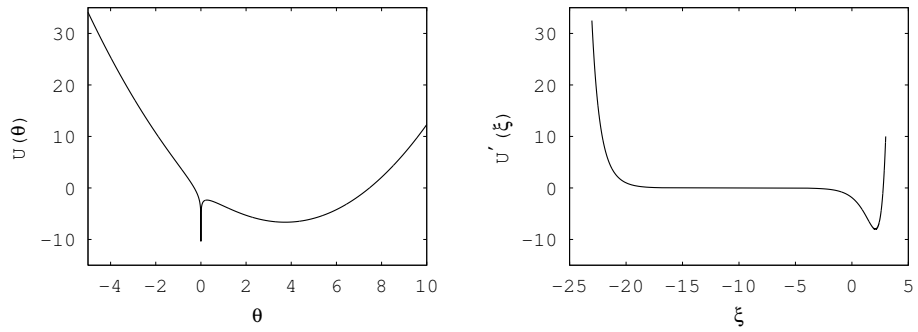
$$\frac{1}{\sqrt{g_i^2 + \alpha}}\mathrm{d}g_i = \mathrm{d}\xi.$$

Integrate both sides to obtain a formula for ξ_i:

$$\xi_i = \log(\theta_i + \sqrt{\theta_i^2 + \alpha}).$$

Then the potential function in the ξ space becomes

$$U'(\xi) = \mathbf{g}(\xi)^{\mathsf{T}}\mathbf{A}\mathbf{g}(\xi) - \mathbf{b}^{\mathsf{T}}\mathbf{g}(\xi),$$



(a) (b)

Figure 5.3. The graphs of $U(\theta)$ (a) and $U'(\xi)$ (b) in the 1-D RVM problem.

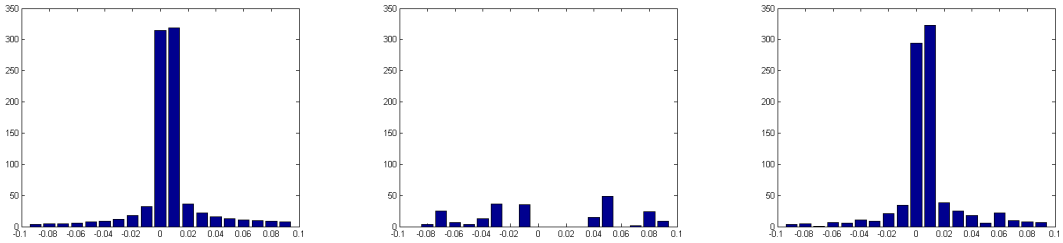
The test problem for this example is a 1-D RVM model. The potential energy function is

$$U(\theta) = A\theta^2 - b\theta + \frac{1}{2}\log(\alpha + \theta^2),$$

with $A = 0.5$, $b = 4$ and $\alpha = 10^{-9}$.

Fig. 5.3 shows the graph of $U(\theta)$ and $U'(\xi)$.

With the same integration step size $\Delta t = 0.5$, both ordinary HMC and HMC with change of variables achieve $> 95\%$ overall acceptance probability. However, ordinary HMC does not sample the region around 0 accurately. The actual acceptance probability in that region is close to 0. Fig. 5.4 shows the number of samples in the region around 0 obtained by ordinary HMC and HMC with change of variables. It can be seen that the histogram obtained by ordinary HMC is not consistent with the desired result, while HMC with change of variables performs well. To be able to sample accuratetly in this region, the integration step size for ordinary HMC must be less than 2.5×10^{-5}. However, this step size is extremely inefficient for the entire region.



(a) (b) (c)

Figure 5.4. Histogram of the number of samples in the region near 0. (a) The target distribution, (b) HMC, and (c) HMCcv

5.2 Two-stage Simplified Takahashi-Imada Method

An improvement in the accuracy of integrating the Hamiltonian system is presented in this section. The leapfrog method, though simple, has only second order accuracy. Fourth order accuracy is achieved by the Takahashi-Imada method [39], which uses the information of the Hessian of the potential energy function. The simplified Takahashi-Imada method [14] avoids the computation of the Hessian. It approximates the Hessian by the evaluation of the force. In [15], the advantage of using a two-stage Hessian based integrator is discussed. Combining the ideas of Refs. [14] and [15], a new fourth order numerical integrator, called the two-stage Takahashi-Imada method is proposed. Experiments are performed on some test problems, and promising results are obtained.

Recall the leapfrog method introduced in Sec. 2.1. It has second order accuracy, i.e., the global error is bounded by $\mathcal{O}(\Delta t^2)$ [14].

The goal here is to find a numerical integrator with higher order accuracy. Fourth order accuracy can be achieved by the Takahashi-Imada method [39], which adds the expression $\Delta t^2(\nabla_\theta \mathbf{f}(\theta))\mathbf{f}(\theta)/12$ to every force evaluation in the leapfrog method for solving the Hamiltonian system, namely

$$\mathbf{p}_{1/2} = \mathbf{p}_0 + \frac{1}{2}(\mathbf{I} + \frac{1}{12}(\Delta t)^2 \nabla_\theta \mathbf{f}(\theta_0))\mathbf{f}(\theta_0)\Delta t. \tag{5.7}$$

Note that this method requires evaluating the Hessian $-\nabla_\theta \mathbf{f}$ of the potential energy function.

To avoid the evaluation of the Hessian, $(\mathbf{I} + (\Delta t)^2 \nabla_\theta \mathbf{f}(\theta)/12)\mathbf{f}(\theta)$ is replaced by $\mathbf{f}(\theta + (\Delta t)^2 \mathbf{f}(\theta)/12)$ to obtain the simplified Takahashi-Imada method [14]:

$$\begin{aligned}
\mathbf{p}_{1/2} &= \mathbf{p}_0 + \frac{1}{2}\mathbf{f}(\theta_0 + \frac{1}{12}(\Delta t)^2 \mathbf{f}(\theta_0))\Delta t, \\
\theta_1 &= \theta_0 + \mathbf{p}_{1/2}\Delta t, \\
\mathbf{p}_1 &= \mathbf{p}_{1/2} + \frac{1}{2}\mathbf{f}(\theta_1 + \frac{1}{12}(\Delta t)^2 \mathbf{f}(\theta_1))\Delta t.
\end{aligned} \tag{5.8}$$

It is of effective order four when combining with the mapping $\chi_{\Delta t} : (\theta, \mathbf{p}) \to (\hat{\theta}, \hat{\mathbf{p}})$ defined by

$$\hat{\theta} = \theta + \frac{1}{12}(\Delta t)^2 \mathbf{f}(\theta),$$
$$\mathbf{p} = \hat{\mathbf{p}} + \frac{1}{12}(\Delta t)^2 \nabla_\theta \mathbf{f}(\theta)\hat{\mathbf{p}}.$$

Let $\Psi_{\Delta t}$ denote the flow of the simplified Takahashi-Imada method (Eq. (5.8)). Effective order four means that the $\chi_{\Delta t} \circ \Psi_{\Delta t} \circ \chi_{\Delta t}^{-1}$ approximates the exact flow $\Phi_{\Delta t}$ to order four.

The two-stage method

$$\begin{aligned}
\mathbf{p}_{1/6} &= \mathbf{p}_0 + \frac{1}{6}\mathbf{f}(\theta_0)\Delta t, \\
\theta_{1/2} &= \theta_0 + \frac{1}{2}\mathbf{p}_{1/6}\Delta t, \\
\mathbf{p}_{5/6} &= \mathbf{p}_{1/6} + \frac{2}{3}(\mathbf{I} + \frac{\Delta t^2}{24}\nabla_\theta \mathbf{f}(\theta_{1/2}))\mathbf{f}(\theta_{1/2})\Delta t, \\
\theta_1 &= \theta_{1/2} + \frac{1}{2}\mathbf{p}_{5/6}\Delta t, \\
\mathbf{p}_1 &= \mathbf{p}_{6/5} + \frac{1}{6}\mathbf{f}(\theta_1)\Delta t.
\end{aligned} \tag{5.9}$$

has fourth order accuracy [15], which does not require any transformation. However, similar to the Takahashi-Imada method, this method also requires evaluating the Hessian.

Therefore, by the same idea of the simplified Takahashi-Imada method, using the first derivative to approximate the second derivative, the two-stage simplified Takahashi-Imada method is obtained by changing the third step of the two-stage method to

$$\mathbf{p}_{5/6} = \mathbf{p}_{1/6} + \frac{2}{3}\mathbf{f}(\theta_{1/2} + \frac{\Delta t^2}{24}\mathbf{f}(\theta_{1/2}))\Delta t. \tag{5.10}$$

The stability condition for the linear test equation with $f(\theta) = -\omega^2\theta$ of both simplified Takahashi-Imada method and two-stage simplified Takahashi-Imada method is $\omega\Delta t < 2\sqrt{3}$. The leapfrog method requires one force evaluation per integration step, the simplified Takahashi-Imada method requires two force evaluations per step,

and the two-stage simplified Takahashi-Imada method requires three force evaluations. However, empirical evidence, given below, shows that the two-stage simplified Takahashi-Imada method is more efficient than the other two methods, when the same number of force evaluations are performed for a given integration interval.

Experiments are performed on two test problems. The first one is an N-dimensional Gaussian problem, which has the potential energy $U(\theta) = \theta^{\mathsf{T}}\theta/2$, Let $\Delta t = 1/3$ for HMC, $\Delta t = 2/3$ for simplified Takahashi-Imada, and $\Delta t = 1$ for two-stage simplified Takahashi-Imada. The integration interval $\tau = 2$ and the number of samples $T = 10^4$. Fig. 5.5 shows the acceptance probability vs. the dimensionality.

The second test problem is again a double-well potential problem similar to the one described in Sec. 2.3.3. The difference is that this test problem has a much higher dimension than that in Sec. 2.3.3. Let $\Delta t = \pi/6$ for HMC, $\Delta t = \pi/3$ for simplified Takahashi-Imada, and $\Delta t = \pi/2$ for two-stage simplified Takahashi-Imada. The integration interval $\tau = \pi$, and the number of samples $T = 10^4$. Fig. 5.6 shows the acceptance probability vs. the dimensionality.

With the chosen Δt and τ, the number of force evaluations for all methods in one Monte Carlo step are all equal to 6. With the same computation cost, it can be seen that the two-stage simplified Takahashi-Imada method has much higher acceptance probability, implying more efficiency, than the other two methods.

5.3 Discussion and Conclusion

The idea of changing variables reduces the difficulty of sampling by transforming the original variable to a new variable. More work needs to be done, however, to make it generally applicable in practice. For example, for the purpose of lowering the energy barrier, a formula of changing variables needs to be found in higher dimensional problems. Techniques for adaptively locating and measuring the shape of the barrier may also be needed. For the two-stage simplified Takahashi-Imada method, preliminary results show the superiority over the leapfrog method. More evidence,

especially that from real applications, are still desired to further justify the utility of the method.

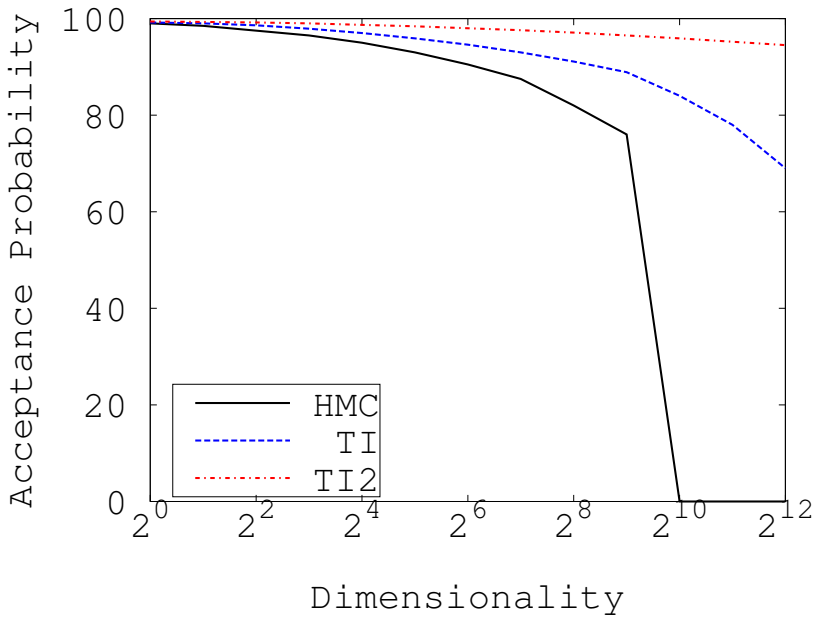


Figure 5.5. Acceptance probability vs. dimensionality for the two-stage simplified Takahashi-Imada method (TI2) and two other methods in the N-dimensional Gaussian problem.

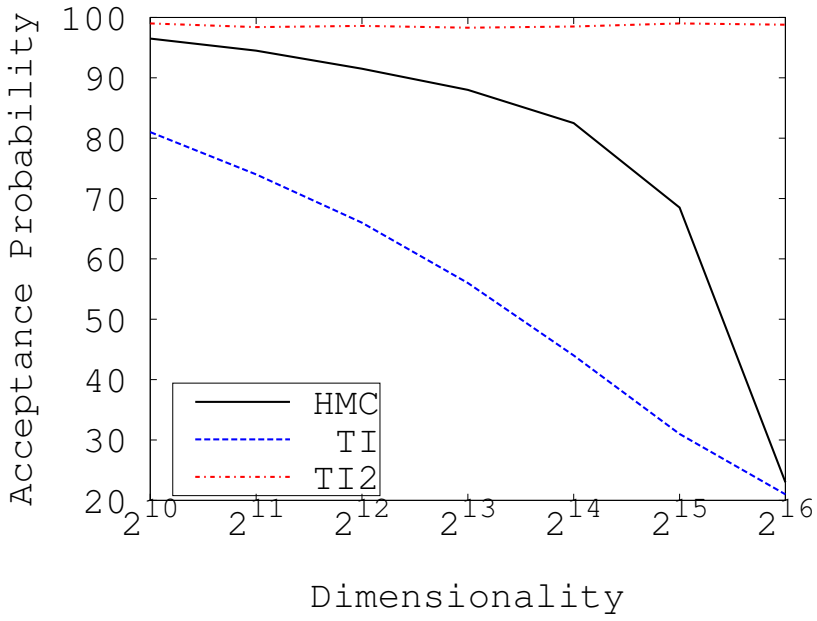


Figure 5.6. Acceptance probability vs. dimensionality for the two-stage simplified Takahashi-Imada method (TI2) and two other methods in the double-well potential problem.

REFERENCES

REFERENCES

[1] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21:1087–1092, 1953.

[2] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physical Review B*, 195:216–222, 1987.

[3] A. M. Horowitz. A Generalized Guided Monte-Carlo Algorithm. *Physics Letters B*, 268:247–252, 1991.

[4] M. Welling and Y. W. Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. *Proceedings of the 28th International Conference on Machine Learning*, 2011.

[5] S. Ahn, A. K. Balan, and M. Welling. Bayesian Posterior Sampling via Stochastic Gradient Fisher Scoring. *Proceedings of the 29th International Conference on Machine Learning*, pages 1591–1598, 2012.

[6] T. Chen, E. B. Fox, and C. Guestrin. Stochastic Gradient Hamiltonian Monte Carlo. *Proceedings of the 31st International Conference on Machine Learning*, 2014.

[7] A. Sokal. Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms. In C. DeWitt-Morette, P. Cartier, and A. Folacci, editors, *Functional Integration: Basics and Applications*, pages 131–192. Springer US, Boston, MA, 1997.

[8] C. J. Geyer. Practical Markov Chain Monte Carlo. *Statistical Science*, 7:473–511, 1992.

[9] P. Minary, M. E. Tuckerman, and G. J. Martyna. Dynamical Spatial Warpping: A Novel Method for the Conformational Sampling of Biophysical Structure. *SIAM Journal on Scientific Computing*, 30(4):2055–2083, 2008.

[10] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, New York, 2004.

[11] Y. Fang, J. M. Sanz-Serna, and R. D. Skeel. Compressible Generalized Hybrid Monte Carlo. *The Journal of Chemical Physics*, 140:174108 (10 pages), 2014.

[12] B. Mehlig, D. W. Heermann, and B. M. Forrest. Hybrid Monte Carlo Method for Condensed-matter Systems. *Physical Review B*, 45:679–685, 1992.

[13] B. Leimkuhler, C. Matthews, and G. Stoltz. The Computation of Averages from Equilibrium and Nonequilibrium Langevin Molecular Dynamics. *IMA Journal of Numerical Analysis*, 36:13–79, 2016.

[14] E. Hairer, R. I. McLachilan, and R. D. Skeel. On Energy Conservation of the Simplified Takahashi-Imada Method. *ESAIM: Mathematical Modelling and Numerical Analysis*, 43:631–644, 2009.

[15] S. Blanes and F. Casas. On the Necessity of Negative Coefficients for Operator Splitting Schemes of Order Higher than Two. *Applied Numerical Mathematics*, 54:23–37, 2005.

[16] R. D. Skeel and Y. Fang. Comparing Markov Chain Samplers for Molecular Simulation. *Entropy*, 10(561), 2017.

[17] V. I. Arnold. *Mathematical Methods of Classical Mechanics*. Springer, 1989.

[18] M. Girolami and B. Calderhead. Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods. *Journal of the Royal Statistical Society: Series B*, 73, Part 2:123–214(with discussion), 2011.

[19] B. Leimkuhler and S. Reich. A Metropolis Adjusted Nosé-Hoover Thermostat. *ESAIM: Mathematical Modelling and Numerical Analysis*, 43(4):743–755, 2009.

[20] M. E. Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation*. Oxford University Press, 2010.

[21] D. J. Evans and G. Morriss. The Isothermal Isobaric Molecular Dynamics Ensemble. *Physics Letters A*, 98(8-9):433–436, 1983.

[22] R. D. Skeel. What Makes Molecular Dynamics Work? *SIAM Journal on Scientific Computing*, 31(2):1363–1378, 2009.

[23] P. Minary, G. J. Martyna, and M. E. Tuckerman. Algorithms and Novel Applications Based on the Isokinetic Ensemble. I. Biophysical and Path Integral Molecular Dynamics. *The Journal of Chemical Physics*, 118:2510, 2003.

[24] C. Schütte and W. Huisinga. Biomolecular Conformations as Metastable Sets of Markov Chain. *Proceedings of the 38th Annual Allerton Conference on Communication, Control and Computing,*, pages 1106–1115, 2000.

[25] G. O. Roberts and R. L. Tweedie. Exponential Convergence of Langevin Diffusions and Their Discrete Approximations. *Bernoulli*, 2:341–363, 1996.

[26] N. Ding, Y. Fang, R. Babbush, C. Chen, R. D. Skeel, and H. Neven. Bayesian Sampling Using Stochastic Gradient Thermostats. In *Advances in Neural Information Processing Systems 27*, 2014.

[27] R. Salakhutdinov and A. Mnih. Bayesian Probabilistic Matrix Factorization Using Markov Chain Monte Carlo. *Proceedings of the 25th International Conference on Machine Learning*, pages 880–887, 2008.

[28] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, March 2003.

[29] S. Patterson and Y. W. Teh. Stochastic Gradient Riemannian Langevin Dynamics on the Probability Simplex. *Advances in Neural Information Processing Systems 26*, pages 3102–3110, 2013.

[30] E. Lyman and D. M. Zuckerman. On the Structural Convergence of Biomolecular Simulations by Determination of the Effective Sample Size. *The Journal of Physical Chemistry B*, 111:12876–12882, 2007.

[31] X. Zhang, D. Bhatt, and D. M. Zuckerman. Automated Sampling Assessment for Molecular Simulations Using the Effective Sample Size. *Journal of Chemical Theory and Computation*, 6:3048–3057, 2010.

[32] J. Goodman. Acor, Statistical Analysis of a Time Series. http://www.math.nyu.edu/faculty/goodman/software/acor/, Spring 2009.

[33] M. B. Priestly. *Spectral Analysis and Time Series*. Academic Press, 1981.

[34] Y. Fang, Y. Cao, and R. D. Skeel. Quasi-reliable Estimates of Effective Sample Size. arXiv:1705.03831, 2017.

[35] H. Risken. *The Fokker-Planck Equation*. Springer, 1996.

[36] C. M. Bishop. *Neural Network for Pattern Recognition*. Clarendon Press, 1996.

[37] M. Lichman. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml, 2013.

[38] M. E. Tipping. Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

[39] M. Takahashi and M. Imada. Monte Carlo Calculation of Quantum Systems. II. Higher Order Correction. *Journal of the Physical Society of Japan*, 53:3765–3769, 1984.