Irreversibility in Stochastic Dynamic Models and Efficient Bayesian Inference

Yian Ma

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Hong Qian, Chair

Emily B. Fox

Steven L. Brunton

Program Authorized to Offer Degree:
Department of Applied Mathematics

University of Washington

Abstract

Irreversibility in Stochastic Dynamic Models and Efficient Bayesian Inference

Yian Ma

Chair of the Supervisory Committee:
Professor Hong Qian
Applied Mathematics

This thesis is the summary of an excursion around the topic of reversibility. We start the journal from a classical mechanical view of the "time reversal symmetry": we look into the details to track the movements of all particles at all times and ask whether the entire system remains the same if both time and momentum flip signs. This description of reversible process is the exact reflection of classical mechanics with a quadratic kinetic energy which generates Boltzmann's equilibrium thermodynamics. Unfortunately, it heavily depends on the coordinate system the variables reside in and automatically excludes the processes with dissipation or/and fluctuation from being reversible. A related but slightly more relaxed scenario is that the dynamics conserve certain quantities. Fortunately, we are able to generalize thermodynamics to this broader range of systems.

For the discussion of reversibility, however, we veer towards a direction that requires much less scrutiny, and provides far more generality. We follow Kolmogorov's footsteps and only study the *statistics* of the variables in question. Reversibility in that realm dictates that the probability of observing a path forward equals to that of seeing a path backward. Interestingly though, the aforementioned conservative dynamics are the source of irreversibility in stationarity. We then realize that the general Markov process can be decomposed into reversible and irreversible components, each preserving the entire process' stationary distribution. This realization lets us continue along the path to develop thermodynamic theory for general stochastic processes and confirm the universal ideal behavior in Orntein-Uhlenbeck processes.

The realization also prompts us to continue our excursion further into applications. On the modeling side, we discover a way to analyze noise induced phenomena in reaction diffusion equations. Stability and bifurcation analysis is brought into the stochastic models through the bridge of "effective dynamics". We are able to quantitatively explain the onset of pattern formations introduced by chemical reaction noise.

Looking over to the Bayesian inference side (for the learning of model parameters from data), we find ourselves in the position of digging into a critical problem: computation with stochasticity. As the defacto approaches for Bayesian inference, Markov chain Monte Carlo (MCMC) methods have always been criticized for their slow convergence (mixing rates) and huge amount of computation required for large data sets (scalability). It has been discovered that introduction of irreversibility increases the mixing of Markov processes. Using the decomposition of general Markov processes, we reparametrize the space of viable Markov processes for sampling purpose, so that the search for the correct MCMC algorithm turns into a game of plug and play with two matrices (or transition probabilities) to choose from. Irreversibility is automatically incorporated as one of the components to specify.

Digging even deeper into a new world of scalable Bayesian inference, we start to make use of stochastic gradient techniques for excessively large data sets. With independent and identically distributed data, our previous results with continuous Markov process can be revised and provide a complete recipe to construct new stochastic gradient MCMC algorithms. Within our recipe, we pick some of the nice attributes of the previous methods and combine them to form an algorithm that excels at learning topics in Wikipedia entries in a streaming manner. With correlated data, we find a huge void space to explore. As the first step, we visit time dependent data and harness the memory decay to generalize the stochastic gradient MCMC methods to hidden Markov models. We find our method about 1,000 times faster than the traditional sampling method for an ion channel recording containing 209,634 observations.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure Number Page

ACKNOWLEDGMENTS

I want to first thank my advisors Hong Qian and Emily Fox for their advice, support, and mentorship. Their help made my success and this thesis possible.

I would also like to thank all the group members in Hong and Emily's groups for all their encouragements, support and accommodating me. Especially, I would like to acknowledge Tianqi Chen, Nick Foti, Felix Ye, Chris Aicher, and Lei Wu. Collaboration with them has been very enjoyable and fruitful.

On a personal note, I want to thank my parents for caring and inspiring me. At last, I want to thank Fan Zhang for always being there for me.

Chapter 1
INTRODUCTION

> Mathematics constitutes the most colossal metaphor imaginable, and must be judged, aesthetically as well as intellectually in terms of the success of this metaphor.
>
> *Norbert Wiener*

Time reversibility has been one of the central concepts for the physics community since the 19th century. Whether a "scientific law" remains the same with the reversal of time is indeed fascinating. At the same time, answer to the above question is contingent upon the very definition of the "scientific laws". Naively, there seem to be two types of reasoning based on different perspectives: In dynamical systems theory, people tend to look for time reversal symmetries and use that notion as the criteria for reversibility. In stochastic processes, time reversibility is often studied from the statistical point of view; reversibility is judged from the statistical indistinguishability between a process and an appropriately defined reversal. In both cases, a type of *invariance* under time reversal is introduced.

In 1948, Norbert Wiener started his book of "Cybernetics" by discussing the difference between the disciplines of astrophysics and atmospheric science [183]. He stated that (in modern language) the astronomical phenomena inherently possess time reversal symmetry; while the atmospheric questions are statistical in nature, and does not have this property of time reversal symmetry. The underlying reasoning behind this distinction is that the astronomical phenomena can be described, with fair accuracy, by Newtonian dynamics of finitely many point masses, which is a closed mechanical system. Reversing time is equivalent to merely reversing the initial momentum of the system. Objects of atmospheric science, on the other hand, are measures over the set of possible atmospheres. The inevitable introduction of stochasticity and dissipation leads to the breaking of time reversal symmetry. One observes that dynamics according to Newtonian mechanics with conservative force, or Hamiltonian dynamics, is in a sense "stationary". Therefore, an appropriate notion of reversibility in the stochastic context has to be based on a stationary stochastic process.

A broader definition of time reversibility, in the theory of stochastic processes, was given by Andrey Kolmogorov [79]. This stochastic processes perspective is based on treatment of observed quantities as stochastic processes, whether they are measurements of celestial objects or collective quantities like clouds. As a direct generalization of the classical, deterministic dynamical systems, a Markov process takes into account the effect of randomness in the motion in a wide range of scenarios. By analyzing the change of a distribution function over time, a statistical description of the quantities of interest can be obtained. A system with an invariant distribution is defined as statitically "stationary". Then it is defined that a Markov process is time reversible if the probability of observing a forward path from one state to another is equal to that of moving backward along the exact same path.

In the following sections, we start by analyzing the notion of time reversal symmetry in the stochastic context and its relationship with conservation law, a notion mainly developed in the context of deterministic dynamical systems. We then study their roles in general stochastic processes with irreversible components.

1.1 Time Reversal Symmetries

Systems following classical mechanics with positions q and momentum p without dissipation have the symmetry: $t \to -t$, $q \to q$, $p \to -p$ leaving the system invariant [88]. In other words, when time runs backwards, the whole system becomes a mirror image of itself in the p (momentum) direction, and hence, leaves the behavior the same. We hasten to add that when a magnetic field is involved, the time reversal symmtry also stipulates that $B \to -B$. The Lorentz force then has a time reversal symmetry $\vec{v} \times \vec{B} \to (-\vec{v}) \times (-\vec{B})$, where $\vec{v}$ is velocity.

In linear systems, time reversal symmetry exists when there is a proper quadratic conserved quantity, and *vice versa.* Systems with a conserved quantity $\phi(\mathbf{x})$ satisfy:

$$f(\mathbf{x})^T \nabla \phi(\mathbf{x}) = 0, \qquad (1.1)$$

which will be discussed in more details below. In nonlinear dynamics, although time reversal symmetry and conservation law do not imply each other, having a proper conserved quantity may lead to other continuous symmetries and provides a recurrence condition.

More importantly, having a proper conserved quantity is the necessary and sufficient condition for constructing a set of "thermodynamic relations" for the dynamical system. The idea of ***thermodynamics*** has not been widely appreciated outside mathematical physics [24]. In an applied mathematical sense, it concerns with the long-time behavior of a recurrent but non-ergodic system, and its parameter dependence, of systems with conserved quantities. Boltzmann and Gibbs first provided the 19^{th} century phenomenological thermophysics with a classical mechanical foundation through its mechanical energy conservation [50]. They realized that the conservation law provides a novel point of view that regards the dynamics on a whole trajectory as a single state of recurrent motion e.g., a *thermodynamic state*. This coarse perspective constitutes a global way of describing and identifying complex dynamical systems, and their parts.

1.1.1 Linear Conservative Flow and General Time Reversal Symmetries

For a linear system with a quadratic conserved quantity $\phi(\mathbf{x}) = \dfrac{1}{2}\mathbf{x}^T\Xi^{-1}\mathbf{x}$, where Ξ^{-1} is symmetric positive definite to ensure that $\phi(\mathbf{x})$ is a proper function in $\mathbb{R}^n$, it can always be written as:

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = Q\Xi^{-1}\mathbf{x}. \tag{1.2}$$

where Q is a skew-symmetric matrix as will be constructively proved in 3.

The dynamics in Eq. 1.2 can be proved as purely cyclic motion (e.g., periodic, or quasi-periodic on an invariant torus). Because matrix QU has only imaginary eigenvalues $\{\lambda_\ell | 1 \leq \ell \leq n\}$ as will be shown in Proof 4. We can also find real Jordan form of QU: PJP^{-1}, where J is block diagonal, with 2×2 skew-symmetric blocks:

$$\mathrm{Im}\left[\lambda_{(2i-1)}\right]\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

being the ith block on the diagonal. Natural coordinates for the conservative flow Eq. (1.2) is therefore: $\mathbf{y} = P^{-1}\mathbf{x}$.

Poisson bracket $\{\cdot,\cdot\}$ can be defined for the linear conservative system as: $\{\varphi(\mathbf{x}), \psi(\mathbf{x})\} = \nabla\varphi(\mathbf{x})^T Q \nabla\psi(\mathbf{x})$. Then the conservative flow expressed in terms of its Hamiltonian function $\phi(\mathbf{x})$ is:

$$\dot{x_i} = \left\{x_i, \tfrac{1}{2}\phi(\mathbf{x})\right\}. \tag{1.3}$$

The conservative flow is totally integrable, with the first integrals I_i:

$$I_i = y_{2i-1}^2 + y_{2i}^2 = \mathbf{x}^T P^{-T} I_{(2i-1)\sim(2i)} P^{-1}\mathbf{x}, \ \ 1 \leq i \leq \left\lfloor \frac{n}{2} \right\rfloor. \tag{1.4}$$

Here, $I_{(2i-1)\sim(2i)}$ denotes the diagonal matrix with 1 on $(2i-1)$-th to $(2i)$-th diagonal entries, and zero everywhere else. There are $\left\lfloor \frac{n}{2} \right\rfloor$ first integrals, but for the given Poisson bracket, one combination of them is unique, which is the Hamiltonian φ that generates the conservative flow.

In the $\mathbf{y}$ coordinates, it is observable that the system bears the following symmetries: $(t, y_{2i-1}, y_{2i}) \longrightarrow \big(-t, (-1)^{k_i} y_{2i-1}, (-1)^{k_i+1} y_{2i}\big)$, where $\{k_i\}$ is a sequence of 0 and 1. Taking $\{k_i\}$ as a sequence of zeros, we recover the time-reversal invariance in classical mechanics. Hence, for general $\{k_i\}$, those symmetries are the natural generalizations of the time-reversal symmetry.

In nonlinear systems, time reversal symmetries and conservation laws do not correspond so well. It can be seen in the next section that conservation laws provide the necessary foundations for a theory of thermodynamics, one that stems from and generalizes the classical mechanical systems with time reversal symmetries.

1.2 Reversibility of Markov Processes

It is interesting to note, on the other hand, that in Kolmogorov's theory of Markov processes, the dynamics with a conserved quantity are what makes the whole process "irreversible" in stationary, while the rest part of the dynamics are found to be time reversible [**?**]. For time homogeneous processes, definition of reversibility according to Kolmogorov is given for a stationary Markov process as

$$p(\mathbf{x}|\mathbf{y})\pi(\mathbf{y}) = p(\mathbf{y}|\mathbf{x})\pi(\mathbf{x}), \tag{1.5}$$

where $p(\mathbf{y}|\mathbf{x})$ is the transition probability from $\mathbf{x}$ to $\mathbf{y}$, and $\pi(\mathbf{x})$ is the stationary probability of $\mathbf{x}$, for any $\mathbf{x}$. In terms of trajectories, the Kolmogorov's criteria asserts that the transition probability must satisfy the following equality that the probability of observing a forward trajectory is equal to the probability of seeing a backward trajectory:

$$p(\mathbf{x}_T|\mathbf{x}_1) \cdot \prod_{i=1}^{T-1} p(\mathbf{x}_{i+1}|\mathbf{x}_i) = p(\mathbf{x}_1|\mathbf{x}_T) \cdot \prod_{i=1}^{T-1} p(\mathbf{x}_i|\mathbf{x}_{i+1}), \tag{1.6}$$

for any finite sequence of random variables $\mathbf{x}_1, \ldots, \mathbf{x}_T$. For nonhomogeneous processes, the Kolmogorov's criteria can be extended to the equality of Crook's and Hatano-Sasa's [33, 64]:

$$\left\langle \frac{\mathcal{P}[\check{\mathbf{x}}(\tau)]}{\mathcal{P}[\mathbf{x}(\tau)]} \right\rangle_{\big[\mathbf{x}(\tau)\big]} = \int \mathcal{D}[\mathbf{x}(\tau)]\ \mathcal{P}[\check{\mathbf{x}}(\tau)] = 1, \tag{1.7}$$

where $\check{\mathbf{x}}(\tau) = \mathbf{x}(T-\tau)$ is the time reversed process of $\mathbf{x}(\tau)$. We will discuss this definition in more depth in Sec. 3.2.

1.3 Implications of Irreversibility

Apart from the fundamental quest, discussion of reversibility incurs greater applied science and engineering consequences. In Chapter 2, we propose a complete framework to decompose general stochastic process into reversible and irreversible components according to its stationary distribution. While obtaining a stationary distribution usually is a challenging task for scientific models; it is actually given *a priori* in any Monte Carlo sampler. In Chapter 3, we will show that the dynamics with a conserved quantity, the irreversible processes, laid the foundation for a general thermodynamic theory of complex dynamics, *à la* Boltzmann and Gibbs. We make use of the irreversible part of stochastic dynamics to rebuild thermodynamics in general stochastic processes. In Chapter 4, we use the decomposition framework to find the most probable effective dynamics of a stochastic reaction diffusion system and analyzing the phenomenon of noise induced pattern formation. In Chapters 5 and 6, we discuss the accelerating effect of irreversibility in Markov chain Monte Carlo sampling algorithms and use our framework to reparameterize Markov dynamics and provide a complete recipe of Markov dynamics to choose from for the design of MCMC methods.

Chapter 2

DECOMPOSITION OF STOCHASTIC PROCESSES ACCORDING TO REVERSIBILITY

2.1 Preliminaries

In this section, we review some of the fundamentals of the stochastic processes associated with general Markov processes.

2.1.1 General Markov Processes and their Diffusion and Jump Operators

Consider a general Markov process in $\mathbb{R}^d$, described by the Chapman-Kolmogorov (CK) equation:

$$p(\mathbf{z};t_3|\mathbf{x};t_1) = \int_{\mathbb{R}^d} \mathbf{d}\mathbf{y} p(\mathbf{z};t_3|\mathbf{y};t_2)p(\mathbf{y};t_2|\mathbf{x};t_1), \tag{2.1}$$

where $t_1 < t_2 < t_3$ are three arbitrary scalar variables denoting time and $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^d$. For succinctness of presentation, in what follows, we write $t = t_2 - t_1$ and let $p(\mathbf{y}|\mathbf{x};t) = p(\mathbf{y};t_2|\mathbf{x};t_1)$ denote the probability of transition from $\mathbf{x}$ to $\mathbf{y}$ over time t and assume autonomous Markov processes (i.e., with time invariant Markov transition operators). It is worth noting that the autonomous assumption is not necessary to any of the calculations; non-autonomous Markov processes bear exactly the same results.

A differential form of (2.1), from which algorithms are more straightforwardly derived, can be obtained by assuming three mild existence conditions for all $\epsilon > 0$:

1. $\lim\limits_{\Delta t \to 0} p(\mathbf{x}|\mathbf{z};\Delta t)/\Delta t = W(\mathbf{x}|\mathbf{z})$ exists uniformly in $\mathbf{x}$ and $\mathbf{z}$ for $|\mathbf{x} - \mathbf{z}| \geq \epsilon$;

2. $\lim\limits_{\Delta t \to 0} \dfrac{1}{\Delta t} \displaystyle\int_{|\mathbf{x}-\mathbf{z}|<\epsilon} \mathbf{d}\mathbf{x}(\mathbf{x}_i - \mathbf{z}_i)p(\mathbf{x}|\mathbf{z};\Delta t) = \mathbf{f}_i(\mathbf{z}) + O(\epsilon)$;

3. $\lim\limits_{\Delta t \to 0} \dfrac{1}{\Delta t} \displaystyle\int_{|\mathbf{x}-\mathbf{z}|<\epsilon} \mathbf{d}\mathbf{x}(\mathbf{x}_i - \mathbf{z}_i)(\mathbf{x}_j - \mathbf{z}_j)p(\mathbf{x}|\mathbf{z};\Delta t) = 2\mathbf{D}_{ij}(\mathbf{z}) + O(\epsilon)$ uniformly in $\mathbf{z}$ and ϵ.

The differential CK equation defines an update rule that consists of a diffusion process and a jump process [51]:

$$\frac{\partial}{\partial t}p(\mathbf{z}|\mathbf{y};t) = \sum_{i,j=1}^{d} \frac{\partial^2}{\partial \mathbf{z}_i \partial \mathbf{z}_j}\big[\mathbf{D}_{ij}(\mathbf{z})p(\mathbf{z}|\mathbf{y};t)\big] - \sum_{i=1}^{d}\frac{\partial}{\partial \mathbf{z}_i}\big[\mathbf{f}_i(\mathbf{z})p(\mathbf{z}|\mathbf{y};t)\big] \\ + \int_{\mathbb{R}^d} \mathbf{dx}\Big[W(\mathbf{z}|\mathbf{x})p(\mathbf{x}|\mathbf{y};t) - W(\mathbf{x}|\mathbf{z})p(\mathbf{z}|\mathbf{y};t)\Big]. \tag{2.2}$$

Here, $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d$, $W : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a transition probability rate function (which defines the transition kernel in (2.10)), and $\mathbf{D}$ is a $d \times d$ positive semidefinite matrix. The first line denotes a continuous Markov process specified by a diffusion operator on $p(\mathbf{z}|\mathbf{y};t)$; the second line is a jump process defined by the transition rate function $W(\mathbf{z}|\mathbf{x})$.

We rewrite (2.2) as

$$\frac{\partial}{\partial t}p(\mathbf{z}|\mathbf{y};t) = \widehat{\mathcal{L}}\left[\frac{p(\mathbf{z}|\mathbf{y};t)}{\pi(\mathbf{z})}\right] + \widehat{\mathcal{J}}\left[\frac{p(\mathbf{z}|\mathbf{y};t)}{\pi(\mathbf{z})}\right], \tag{2.3}$$

where $\widehat{\mathcal{L}}[\cdot]$ is the diffusion operator defined as:

$$\widehat{\mathcal{L}}\left[\varphi(\mathbf{z})\right] = \sum_{i,j=1}^{d} \frac{\partial^2}{\partial \mathbf{z}_i \partial \mathbf{z}_j}\left[\mathbf{D}_{ij}(\mathbf{z})\pi(\mathbf{z})\varphi(\mathbf{z})\right] - \sum_{i=1}^{n}\frac{\partial}{\partial \mathbf{z}_i}\left[\mathbf{f}_i(\mathbf{z})\pi(\mathbf{z})\varphi(\mathbf{z})\right], \tag{2.4}$$

and $\widehat{\mathcal{J}}[\cdot]$ is a Markov transition operator with kernel $W(\mathbf{z}|\mathbf{x})$:

$$\widehat{\mathcal{J}}\left[\varphi(\mathbf{z})\right] = \int_{\mathbb{R}^d} \mathbf{dx}\left[W(\mathbf{z}|\mathbf{x})\pi(\mathbf{x})\varphi(\mathbf{x}) - W(\mathbf{x}|\mathbf{z})\pi(\mathbf{z})\varphi(\mathbf{z})\right]. \tag{2.5}$$

Equation:

$$\frac{\partial}{\partial t}p(\mathbf{z},t) = \widehat{\mathcal{L}}\left[\frac{p(\mathbf{z},t)}{\pi(\mathbf{z})}\right] \tag{2.6}$$

is called the Fokker-Planck equation; and equation:

$$\frac{\partial}{\partial t}p(\mathbf{z},t) = \widehat{\mathcal{J}}\left[\frac{p(\mathbf{z},t)}{\pi(\mathbf{z})}\right] \tag{2.7}$$

is called Markov jump process.

This form allows us to do two things straightforwardly: One is to separate our analyses of the continuous and jump parts; the second is to analyze the reversibility of the processes. In particular, reversible processes satisfy the algebraic condition that $\dfrac{p(\mathbf{x}|\mathbf{y};t)}{\pi(\mathbf{x})} = \dfrac{p(\mathbf{y}|\mathbf{x};t)}{\pi(\mathbf{y})}$, or as is more commonly written, $p(\mathbf{x}|\mathbf{y};t)\pi(\mathbf{y}) = p(\mathbf{y}|\mathbf{x};t)\pi(\mathbf{x})$. When the operators $\widehat{\mathcal{L}}$ and $\widehat{\mathcal{J}}$ on $p(\mathbf{z}|\mathbf{y};t)/\pi(\mathbf{z})$ are *self-adjoint* (i.e., their adjoint operators in the Hilbert space L^2 are equal to themselves), the Markov process is reversible. Hence, expressing the evolution of the current probability distribution with respect to the stationary distribution as in (2.3) can help reveal this structure.

When the stationary distribution π is not strictly positive or when the space for $\mathbf{z}$ is not compact (e.g., the space $\mathbb{R}^d$ used here), a proper space for $\dfrac{p(\mathbf{z}|\mathbf{y};t)}{\pi(\mathbf{z})}$ may not be obvious. A formal way to express the same idea is through the generator of the diffusion process (corresponding to the Kolmogorov backward equation), where the functions exist in a new Hilbert space L^2_π with inner product defined as $\langle f(\mathbf{x}), g(\mathbf{x})\rangle_\pi = \int f(\mathbf{x})g(\mathbf{x})\pi(\mathbf{x})\mathrm{d}\mathbf{x}$. Then self-adjointness of the generator is equivalent to the reversibility of the continuous Markov process.

2.1.2 Realization of Markov Processes

In Section 2.1.1, we described the evolution of the distribution $p(\mathbf{z}|\mathbf{y};t)$. This evolution provides insight into the stationary distribution of the process. Here, we present the dynamics for an individual realization, which will play a critical role in our developed samplers.

One can generate a stochastic process based on another, usually more elementary, s-tochastic process. In computation Monte Carlo sampling, this is based on the i.i.d. random numbers provided by a computer: for a continuous-state, continuous-time Markov process, it is conveniently based on the standard Brownian motion; and for a discrete-state, continuous-time jump process, it can be based on the standard Poisson process. After simulating (usually approximating) the dynamics, a realization of the Markov process can be obtained.

More specifically, a realization of the continuous Markov process, $\dfrac{\partial}{\partial t}p(\mathbf{z}|\mathbf{y};t) = \widehat{\mathcal{L}}\left[\dfrac{p(\mathbf{z}|\mathbf{y};t)}{\pi(\mathbf{z})}\right]$, can be generated from the stochastic differential equation (SDE) over a real-valued random variable $\mathbf{Z} : \Omega \to \mathbb{R}^d$ from the probability space $(\Omega, \mathcal{F}, P)$ to the measurable space $(\mathbb{R}^d, \mathfrak{B}(\mathbb{R}^d))$ with Borel algebra:

$$\mathrm{d}\mathbf{Z} = \mathbf{f}(\mathbf{Z})\mathrm{d}t + \sqrt{2\mathbf{D}(\mathbf{Z})}\mathrm{d}\mathbf{W}(t), \tag{2.8}$$

where $\sqrt{\mathbf{D}(\mathbf{Z})}$ is defined as a solution to $\mathfrak{D}(\mathbf{Z})$ such that: $\mathfrak{D}(\mathbf{Z})\mathfrak{D}(\mathbf{Z})^{\mathrm{T}} = \mathbf{D}(\mathbf{Z})$ (which always exists and is real for symmetric positive semi-definite $\mathbf{D}(\mathbf{Z})$). In practice, to simulate from (2.8), we consider an ϵ-discretization:

$$\mathbf{Z}_{t+1} \leftarrow \mathbf{Z}_t + \epsilon_t\mathbf{f}(\mathbf{Z}_t) + \eta_t, \quad \eta_t \sim \mathcal{N}(0, 2\epsilon_t\mathbf{D}(\mathbf{Z}_t)) \tag{2.9}$$

Although (2.9) is in the form of the Euler–Maruyama method, higher order numerical schemes can be used for better accuracy [26, 20, 94].

Turning to the Markov jump process, the equation $\frac{\partial}{\partial t}p(\mathbf{z}|\mathbf{y};t) = \widehat{\mathcal{J}}\left[\frac{p(\mathbf{z}|\mathbf{y};t)}{\pi(\mathbf{z})}\right]$ is approximately to the first order in Δt given by:

$$p(\mathbf{z}|\mathbf{y};t+\Delta t) = \Delta t\int_{\mathbb{R}^d} W(\mathbf{z}|\mathbf{x})p(\mathbf{x}|\mathbf{y};t)\mathrm{d}\mathbf{x} + \left[1-\Delta t\int_{\mathbb{R}^d} W(\mathbf{x}|\mathbf{z})\mathrm{d}\mathbf{x}\right]p(\mathbf{z}|\mathbf{y};t).$$

Although this is an approximation to the original equation, it has the same stationary distribution. Noting that $p(\mathbf{x}|\mathbf{y};0) = \delta(\mathbf{z}-\mathbf{y})$, we arise at the transition probability:

$$p(\mathbf{z}|\mathbf{y};\Delta t) = \Delta t W(\mathbf{z}|\mathbf{y}) + \left[1-\Delta t\int_{\mathbb{R}^d} W(\mathbf{x}|\mathbf{y})\mathrm{d}\mathbf{x}\right]\delta(\mathbf{z}-\mathbf{y}). \tag{2.10}$$

Equation (2.10) corresponds to a sampling process as follows. We take Δt to be the stepsize. Then, with probability $\Delta t W(\mathbf{z}|\mathbf{y})$, we transit from state $\mathbf{y}$ to state $\mathbf{z}$. With probability $1-\Delta t\int_{\mathbb{R}^d} W(\mathbf{x}|\mathbf{y})\mathrm{d}\mathbf{x}$, we stay in state $\mathbf{y}$. For fixed Δt, Eq. (2.10) is equivalent to a τ-leaping simulation (revised from the Gillespie algorithm) of a Poisson process.

2.2 Idea from Freidlin Wentzell Theory

As described in Sec. 5.1.1, continuous Markov dynamics can be written as a set of stochastic differential equation (SDE) for the random variable $\mathbf{Z}$:

$$\mathrm{d}\mathbf{Z} = \mathbf{f}(\mathbf{Z})\mathrm{d}t + \sqrt{2\epsilon\mathbf{D}(\mathbf{Z})}\mathrm{d}\mathbf{W}(t). \tag{2.11}$$

Here we let the dynamics to additionally depend on a small noise parameter ϵ. Under Itô's interpretation, the SDE above describes the time evolution of a probability density function $p(\mathbf{z},t)$ over the space that $\mathbf{Z}$ can take value $\mathbf{z}$ according to the Fokker-Planck equation below [146]:

$$\frac{\partial p(\mathbf{z},t)}{\partial t} = \epsilon\sum_{i,j}\frac{\partial^2}{\partial\mathbf{z}_i\partial\mathbf{z}_j}\big(\mathbf{D}_{ij}(\mathbf{z})p(\mathbf{z},t)\big) - \sum_i\frac{\partial}{\partial\mathbf{z}_i}\big(\mathbf{f}_i(\mathbf{z})p(\mathbf{z},t)\big). \tag{2.12}$$

When the stochasticity controlled by ϵ is small, we can use the WKB ansatz to focus on the most probable behavior of the system by assuming [137]:

$$p(\mathbf{z},t) = e^{-\dfrac{\phi(\mathbf{z},t)}{\epsilon} + \phi_0(\mathbf{z},t) + \epsilon\phi_1(\mathbf{z},t) + \mathcal{O}(\epsilon^2)}. \tag{2.13}$$

According to Freidlin-Wentzell theory [48], function $\phi(\mathbf{z}, t)$ on the $\frac{1}{\epsilon}$ order characterizes the behavior around the most probable path [98]:

$$\begin{aligned}\phi(\mathbf{z}, t) &= -\lim_{\epsilon\to 0} \epsilon \ln p(\mathbf{z}, t) \\ &= -\lim_{\epsilon\to 0} \epsilon \left[\ln \int \mathcal{D}\mathbf{z}(t) \exp(-\epsilon^{-1} S_t[\mathbf{z}(t)]) - \ln Z\right] = \inf_{\mathbf{z}(0)=\mathbf{z}_0, \mathbf{z}(t)=\mathbf{z}} S_t[\mathbf{z}(t)],\end{aligned} \tag{2.14}$$

where for a given regular connecting path $\mathbf{z}(t)$ and ϵ, δ small enough,

$$\mathcal{P}(\sup_{0\leq t\leq T} |\mathbf{Z}(t) - \mathbf{z}(t)| \leq \delta) \approx \exp(-\epsilon^{-1} S_T[\mathbf{z}(t)]).$$

And $S_T[\mathbf{z}(t)]$ is called the large deviation rate functional.

Substituting the WKB ansatz of Eq. (2.13) into Eq. (2.12), we obtain the Hamilton-Jacobi equation:

$$\frac{\partial \phi(\mathbf{z}, t)}{\partial t} = -\nabla\phi(\mathbf{z}, t)^{\mathrm{T}} \big(\mathbf{D}(\mathbf{z})\nabla\phi(\mathbf{z}, t) + \mathbf{f}(\mathbf{z})\big). \tag{2.15}$$

Since the Hamilton-Jacobi equation characterizes the the behavior of the stochastic process around its most probable path, it is independence of how Eq. (6.2) is interpreted, whether we use Itô's or Stratonovich's or other types' of stochastic integration. When a nontrivial differentiable stationary solution $\phi^s(\mathbf{z})$ for Eq. (2.15) exists, a conservation law is naturally implied:

$$-\nabla\phi^s(\mathbf{z})^{\mathrm{T}} \big(\mathbf{D}(\mathbf{z})\nabla\phi^s(\mathbf{z}) + \mathbf{f}(\mathbf{z})\big) = 0. \tag{2.16}$$

When $\mathbf{f}(\mathbf{z})$ and $\nabla\phi^s(\mathbf{z})$ have the same zeros, we can write

$$\mathbf{Q}^\phi(\mathbf{z})\nabla\phi^s(\mathbf{z}) = -\big(\mathbf{D}(\mathbf{z})\nabla\phi^s(\mathbf{z}, t) + \mathbf{f}(\mathbf{z})\big), \tag{2.17}$$

where $\mathbf{Q}^\phi(\mathbf{z})$ is skew-symmetric. For example, $\mathbf{Q}^\phi(\mathbf{z})$ can be constructed as [187]:

$$\mathbf{Q}^\phi(\mathbf{z}) = \frac{\mathbf{v}(\mathbf{z})\nabla\phi^s(\mathbf{z})^T - \nabla\phi^s(\mathbf{z})\mathbf{v}(\mathbf{z})^T}{\nabla\phi^s(\mathbf{z})^{\mathrm{T}}\nabla\phi^s(\mathbf{z})}, \quad \mathbf{v}(\mathbf{z}) = -\big(\mathbf{D}(\mathbf{z})\nabla\phi^s(\mathbf{z}, t) + \mathbf{f}(\mathbf{z})\big). \tag{2.18}$$

Hence we decompose $\mathbf{f}$ as:

$$\mathbf{f}(\mathbf{z}) = -(\mathbf{D}(\mathbf{z}) + \mathbf{Q}^\phi(\mathbf{z}))\nabla\phi^s(\mathbf{z}). \tag{2.19}$$

It is straightforward to see that $\phi^s(\mathbf{z})$ is a Lyapunov function for the vector field $\mathbf{f}$. Then Eq. (2.15) becomes:

$$\begin{aligned}\frac{\partial\phi(\mathbf{z},t)}{\partial t} &= -\nabla\phi(\mathbf{z},t)^{\mathrm{T}}\mathbf{D}(\mathbf{z})\nabla(\phi(\mathbf{z},t)-\phi^s(\mathbf{z}))+\nabla\phi(\mathbf{z},t)^{\mathrm{T}}\mathbf{Q}^{\phi}(\mathbf{z})\nabla\phi^s(\mathbf{z}) \\ &= -\nabla\phi(\mathbf{z},t)^{\mathrm{T}}(\mathbf{D}(\mathbf{z})+\mathbf{Q}^{\phi}(\mathbf{z}))\nabla(\phi(\mathbf{z},t)-\phi^s(\mathbf{z})). \qquad (2.20)\end{aligned}$$

On the leading order, Eq. (6.2) has most probable behavior according to the following SDE:

$$\mathrm{d}\mathbf{Z} = -(\mathbf{D}(\mathbf{Z})+\mathbf{Q}^{\phi}(\mathbf{Z}))\nabla\phi^s(\mathbf{Z})\mathrm{d}t+\sqrt{2\epsilon\mathbf{D}(\mathbf{Z})}\mathrm{d}\mathbf{W}(t). \qquad (2.21)$$

It can be seen that Eq. (2.21) connects the stochastic dynamics with its stationary behavior and separates into two parts. One part consists of reversible dynamics:

$$\mathrm{d}\mathbf{Z} = -\mathbf{D}(\mathbf{Z})\nabla\phi^s(\mathbf{Z})\mathrm{d}t+\sqrt{2\epsilon\mathbf{D}(\mathbf{Z})}\mathrm{d}\mathbf{W}(t). \qquad (2.22)$$

It settles at the stationary probability density function $p^s(\mathbf{z}) \propto e^{-\phi^s(\mathbf{z})/\epsilon+\phi_0^s(\mathbf{z})+\epsilon\phi_1^s(\mathbf{z})+\mathcal{O}(\epsilon^2)}$ dominated by $\phi(\mathbf{z})$, and is in a constant state of detailed balance (everything that flows out has the same probability of flowing in from the same direction, nothing really changes on the leading $\mathcal{O}(1)$ order: $p(\mathbf{x},t|\mathbf{y})p^s(\mathbf{y}) = p(\mathbf{y},t|\mathbf{x})p^s(\mathbf{x})$).

Another part is an irreversible conservative motion:

$$\mathrm{d}\mathbf{z} = -\mathbf{Q}^{\phi}(\mathbf{z})\nabla\phi^s(\mathbf{z})\mathrm{d}t, \qquad (2.23)$$

where $\phi^s(\mathbf{z})$ is the conserved quantity of Eq. (2.23).

It can be seen that this dominant behavior of the stochastic process has a general conservation law in stationarity. The incessant but stationary behavior is still driven by a deterministic dynamics, Eq. (2.23), with a conserved quantity $\phi^s(\mathbf{z})$.

2.3 Decomposition Away from Most Probable Path with Itô's Interpretation

Eq. (2.21) gives a nice form of decomposing the dynamics into a reversible and irreversible parts. We wish to obtain similar form of the decomposition for finite noise, at the same time preserve the stationary distribution away from most probable path. This implicitly requires a resummation of all the terms in Eq. (2.13):

$$p(\mathbf{z},t) = e^{-\dfrac{\varphi(\mathbf{z},t)}{\epsilon}}, \quad \varphi(\mathbf{z},t) = \phi(\mathbf{z},t)-\epsilon\phi_0(\mathbf{z},t)-\epsilon^2\phi_1(\mathbf{z},t)+\mathcal{O}(\epsilon^3). \qquad (2.24)$$

After the resummation, an equation similar to the reformulated Hamilton-Jacobi equation, Eq. (2.20), can be written as:

$$\begin{aligned}\frac{\partial\varphi(\mathbf{z},t)}{\partial t} =& -\nabla\varphi(\mathbf{z},t)^{\mathrm{T}}(\mathbf{D}(\mathbf{z})+\mathbf{Q}(\mathbf{z}))\nabla(\varphi(\mathbf{z},t)-\varphi^s(\mathbf{z})) \\ &+\epsilon\nabla^{\mathrm{T}}\Big((\mathbf{D}(\mathbf{z})+\mathbf{Q}(\mathbf{z}))\nabla(\varphi(\mathbf{z},t)-\varphi^s(\mathbf{z}))\Big),\end{aligned} \tag{2.25}$$

where in general, $\mathbf{Q}(\mathbf{z}) \neq \mathbf{Q}^{\phi}(\mathbf{z})$. To parallel with Eq. (2.26), $\mathbf{Q}(\mathbf{z})$ also has a relationship with $\mathbf{f}(\mathbf{z})$, $\mathbf{D}(\mathbf{z})$, and $\varphi^s(\mathbf{z})$:

$$\mathbf{f}(\mathbf{z}) = -(\mathbf{D}(\mathbf{z})+\mathbf{Q}(\mathbf{z}))\nabla\varphi^s(\mathbf{z})+\epsilon\Gamma(\mathbf{z}), \quad \Gamma_i(\mathbf{z}) = \sum_{j=1}^{d}\frac{\partial}{\partial x_j}(\mathbf{D}_{ij}(\mathbf{z})+\mathbf{Q}_{ij}(\mathbf{z})). \tag{2.26}$$

A few comments are in order here.

- First, for finite ϵ, $\varphi^s(\mathbf{z})$ in general has better regularity than $\phi^s(\mathbf{z})$. This is due to the correspondence of $\varphi(\mathbf{z},t)$ and $\varphi^s(\mathbf{z})$ to the generator of diffusion process in Eq. (2.12), and in turn imparted by the nice properties of its resolvent operators. On the other hand, nontrivial stationary solution $\phi^s(\mathbf{z})$ to Eq. 2.15 exists if large deviation principle of Freidlin Wentzell type is satisfied and Eq. (2.16) has nontrivial solution. Even so, solutions to Eq. (2.16) are in general not unique and not smooth. Hence there are far less constraints to work with $\varphi(\mathbf{z},t)$ and $\varphi^s(\mathbf{z})$ than with $\phi(\mathbf{z},t)$ and $\phi^s(\mathbf{z})$.

- Second, since we are no longer just focusing on the leading order behavior around the most probable path, and the stationary probability density function of Eq. (2.29) is exactly: $p^s(\mathbf{z}) \propto e^{-\varphi^s(\mathbf{z})/\epsilon}$, the results are no longer independent of the way the SDEs are interpreted. In this section, we take Itô's interpretation as in the rest of the paper. Then a decomposition of the SDE Eq. (6.2) exists and has a correction term $\Gamma(\mathbf{Z})$ on the order of ϵ in addition to Eq. (2.21):

$$\mathrm{d}\mathbf{Z} = -\big(\mathbf{D}(\mathbf{Z})+\mathbf{Q}(\mathbf{Z})\big)\nabla\varphi^s(\mathbf{Z})\mathrm{d}t+\epsilon\Gamma(\mathbf{Z})\mathrm{d}t+\sqrt{2\epsilon\mathbf{D}(\mathbf{Z})}\mathrm{d}\mathbf{W}(t), \tag{2.27}$$

- Third, it might be tempting to think that Eq. (2.25) is an expansion of the Fokker-Planck equation, Eq. (2.12), to one higher order. However, it is worth noting that $\varphi(\mathbf{z},t)$, $\varphi^s(\mathbf{z})$, and $\mathbf{Q}(\mathbf{z})$ actually contains higher order terms of ϵ. As a result, Eq. (2.25) is actually *exact*. Therefore, one can simply take $\epsilon = 1$ and obtain:

$$\mathrm{d}\mathbf{Z} = -\big(\mathbf{D}(\mathbf{Z})+\mathbf{Q}(\mathbf{Z})\big)\nabla\varphi^s(\mathbf{Z})\mathrm{d}t+\Gamma(\mathbf{Z})\mathrm{d}t+\sqrt{2\mathbf{D}(\mathbf{Z})}\mathrm{d}\mathbf{W}(t). \tag{2.28}$$

- Fourth, the existence of $\mathbf{Q}(\mathbf{z})$ given $\mathbf{f}(\mathbf{z})$, $\mathbf{D}(\mathbf{z})$, and $\varphi^s(\mathbf{z})$ is proved in Theorem 1 below.

In other words, Eq. (2.8) can be generally transformed into the following form [102]:

$$\mathrm{d}\mathbf{Z} = -\big(\mathbf{D}(\mathbf{Z}) + \mathbf{Q}(\mathbf{Z})\big)\nabla\varphi^s(\mathbf{Z})\mathrm{d}t + \Gamma(\mathbf{Z})\mathrm{d}t + \sqrt{2\mathbf{D}(\mathbf{Z})}\mathrm{d}\mathbf{W}(t), \tag{2.29}$$

where the correction term $\Gamma(\mathbf{Z})$ is:

$$\Gamma_i(\mathbf{Z}) = \sum_{j=1}^{d} \frac{\partial}{\partial \mathbf{Z}_j}\big(\mathbf{D}_{ij}(\mathbf{Z}) + \mathbf{Q}_{ij}(\mathbf{Z})\big). \tag{2.30}$$

And its Fokker-Planck equation is equivalent to the following form:

$$\frac{\partial}{\partial t}p(\mathbf{z};t) = \nabla^T \cdot \Bigg(\big[\mathbf{D}(\mathbf{z}) + \mathbf{Q}(\mathbf{z})\big]\big[p(\mathbf{z};t)\nabla\varphi^s(\mathbf{z}) + \nabla p(\mathbf{z};t)\big]\Bigg). \tag{2.31}$$

Theorem 1 *For the SDE of Eq.* (6.2)*, suppose its stationary probability density function* $p^s(\mathbf{z})$ *uniquely exists, and that* $\left[\mathbf{f}_i(\mathbf{z})p^s(\mathbf{z}) - \sum_{j=1}^n \frac{\partial}{\partial \mathbf{z}_j}\Big(\mathbf{D}_{ij}(\mathbf{z})p^s(\mathbf{z})\Big)\right]$ *is integrable with respect to the Lebesgue measure, then there exists a skew-symmetric* $\mathbf{Q}(\mathbf{z})$ *such that Eq.* (2.30) *is equivalent to Eq.* (6.2).

Proof 1 *Comparing Eq.* (2.30) *with Eq.* (6.2)*, we find that the condition for them to be equivalent is:*

$$\mathbf{f}(\mathbf{z}) = -\big[\mathbf{D}(\mathbf{z}) + \mathbf{Q}(\mathbf{z})\big]\nabla\varphi^s(\mathbf{z}) + \Gamma(\mathbf{z}), \tag{2.32}$$

for any $\mathbf{z} \in \mathcal{D}$*. Multiplying* $p^s(\mathbf{z})$ *on both sides of Eq. (6.3), and noting that:*

$$p^s(\mathbf{z}) \propto \exp\left(-\varphi^s(\mathbf{z})\right), \tag{2.33}$$

we arrive at:

$$\mathbf{f}_i(\mathbf{z})p^s(\mathbf{z}) = \sum_j \frac{\partial}{\partial \mathbf{z}_j}\Big(\big(\mathbf{D}_{ij}(\mathbf{z}) + \mathbf{Q}_{ij}(\mathbf{z})\big)p^s(\mathbf{z})\Big). \tag{2.34}$$

The equation for $\mathbf{Q}_{ij}(\mathbf{z})$ *can now be written as:*

$$\sum_j \frac{\partial}{\partial \mathbf{z}_j}\Big(\mathbf{Q}_{ij}(\mathbf{z})p^s(\mathbf{z})\Big) = \mathbf{f}_i(\mathbf{z})p^s(\mathbf{z}) - \sum_j \frac{\partial}{\partial \mathbf{z}_j}\Big(\mathbf{D}_{ij}(\mathbf{z})p^s(\mathbf{z})\Big). \tag{2.35}$$

Recall that the Fokker-Planck equation for the stochastic process, Eq. (6.2), is:

$$\begin{aligned}\frac{\partial p(\mathbf{z},t)}{\partial t} &= -\nabla^T \cdot \big(\mathbf{f}(\mathbf{z})p(\mathbf{z},t)\big) + \nabla^2 : \big(\mathbf{D}(\mathbf{z})p(\mathbf{z},t)\big) \\ &= -\sum_i \frac{\partial}{\partial \mathbf{z}_i}\left\{\mathbf{f}_i(\mathbf{z})p(\mathbf{z},t) - \sum_j \frac{\partial}{\partial \mathbf{z}_j}\Big(\mathbf{D}_{ij}(\mathbf{z})p(\mathbf{z},t)\Big)\right\}. \end{aligned} \tag{2.36}$$

We can immediately observe that the right hand side of Eq. (2.35) has a divergenceless property by substituting the stationary probability density function $p^s(\mathbf{z})$ into Eq. (2.36):

$$\sum_i \frac{\partial}{\partial \mathbf{z}_i}\left\{\mathbf{f}_i(\mathbf{z})p^s(\mathbf{z}) - \sum_j \frac{\partial}{\partial \mathbf{z}_j}\Big(\mathbf{D}_{ij}(\mathbf{z})p^s(\mathbf{z})\Big)\right\} = 0. \tag{2.37}$$

The nice forms of Eqs. (2.35) and (2.37) imply that the questions can be transformed into a linear algebra problem once we apply a Fourier transform to them. Denote the Fourier transform of $\mathbf{Q}(\mathbf{z})p^s(\mathbf{z})$ as $\hat{\mathbf{Q}}(\mathbf{k})$; and Fourier transform of $\mathbf{f}_i(\mathbf{z})p^s(\mathbf{z}) - \sum\limits_j \frac{\partial}{\partial \mathbf{z}_j}\Big(\mathbf{D}_{ij}(\mathbf{z})p^s(\mathbf{z})\Big)$ as $\hat{\mathbf{F}}_i(\mathbf{k})$, where $\mathbf{k} = (\mathbf{k}_1, \cdots, \mathbf{k}_n)^T$ is the set of the spectral variables. That is:

$$\begin{aligned}\hat{\mathbf{Q}}_{ij}(\mathbf{k}) &= \int_{\mathcal{D}} \mathbf{Q}_{ij}(\mathbf{z})p^s(\mathbf{z})e^{-2\pi \mathrm{i}\ \mathbf{k}^T\mathbf{z}}d\mathbf{z}; \\ \hat{\mathbf{F}}_i(\mathbf{k}) &= \int_{\mathcal{D}} \left(\mathbf{f}_i(\mathbf{z})p^s(\mathbf{z}) - \sum_j \frac{\partial}{\partial \mathbf{z}_j}\Big(\mathbf{D}_{ij}(\mathbf{z})p^s(\mathbf{z})\Big)\right) e^{-2\pi \mathrm{i}\ \mathbf{k}^T\mathbf{z}}d\mathbf{z}.\end{aligned}$$

Then, $\frac{\partial}{\partial \mathbf{z}_j}\Big(\mathbf{Q}_{ij}(\mathbf{z})p^s(\mathbf{z})\Big)$ is transformed to $2\pi\mathrm{i}\ \hat{\mathbf{Q}}_{ij}\mathbf{k}_j$, and Eq. (2.35) becomes the following equivalent form in Fourier space:

$$\begin{cases} 2\pi\mathrm{i}\ \hat{\mathbf{Q}}\mathbf{k} = \hat{\mathbf{F}} \\ \mathbf{k}^T\hat{\mathbf{F}} = 0. \end{cases} \tag{2.38}$$

Hence, it is clear that matrix $\hat{\mathbf{Q}}$ must be a skew-symmetric projection matrix from the span of $\mathbf{k}$ to the span of $\hat{\mathbf{F}}$, where $\mathbf{k}$ and $\hat{\mathbf{F}}$ are always orthogonal to each other. We thereby construct $\hat{\mathbf{Q}}$ as combination of two rank 1 projection matrices:

$$\hat{\mathbf{Q}} = (2\pi\mathrm{i})^{-1}\frac{\hat{\mathbf{F}}\mathbf{k}^T}{\mathbf{k}^T\mathbf{k}} - (2\pi\mathrm{i})^{-1}\frac{\mathbf{k}\hat{\mathbf{F}}^T}{\mathbf{k}^T\mathbf{k}}. \tag{2.39}$$

We arrive at the final result that matrix $\mathbf{Q}(\mathbf{z})$ is equal to $p^s(\mathbf{z})^{-1}$ times the inverse Fourier transform of $\hat{\mathbf{Q}}(\mathbf{k})$:

$$\mathbf{Q}_{ij}(\mathbf{z}) = p^s(\mathbf{z})^{-1}\int_{\mathcal{D}} \frac{\mathbf{k}_j\hat{\mathbf{F}}_i(\mathbf{k}) - \mathbf{k}_i\hat{\mathbf{F}}_j(\mathbf{k})}{(2\pi\mathrm{i})\cdot\sum\limits_l \mathbf{k}_l^2} e^{2\pi\mathrm{i}\sum\limits_l \mathbf{k}_l\mathbf{z}_l} d\mathbf{k}. \tag{2.40}$$

Thus, if $\left(\mathbf{f}_i(\mathbf{z})p^s(\mathbf{z}) - \sum_j \frac{\partial}{\partial \mathbf{z}_j}\Big(\mathbf{D}_{ij}(\mathbf{z})p^s(\mathbf{z})\Big)\right)$ *belongs to the space of* L^1*, then any continuous time Markov process, Eq. (6.2), can be turned into this new formulation.*

To the best of our knowledge, the exact form of Eq. (2.29) was first presented in the statistical mechanics literature [186, 162]; however, the completeness of the representation of continuous Markov processes was made only later in [102]. The proof of Theorem 1 is comprised of two sets of ideas stemming from different fields: In studies of continuous Markov processes, earlier works [139, 34, 134, 35, 175, 123] realized that diffusion processes with stationary probability density function $\pi(\mathbf{z})$ can be decomposed into a reversible part and an irreversible part which preserves $\pi(\mathbf{z})$ as its invariant measure. In stochastic models in fluid dynamics and homogenization, earlier works [81] found that divergenceless vector fields can be represented as the divergence of an anti-symmetric matrix valued potential. Combination of both ideas can lead to the discovery of Eq. (2.31) that underlies the proof of Theorem 1. Similar structures have also been seen in even earlier works when one or both of $\mathbf{D}(\mathbf{z})$ and $\mathbf{Q}(\mathbf{z})$ are constant matrices [70, 87].

An important feature of this decomposition framework in Eq. (2.29) is that it builds a bridge between the stochastic differential equation:

$$\mathrm{d}\mathbf{Z} = -\big(\mathbf{D}(\mathbf{Z}) + \mathbf{Q}(\mathbf{Z})\big)\nabla\varphi^s(\mathbf{Z})\mathrm{d}t + \Gamma(\mathbf{Z})\mathrm{d}t + \sqrt{2\mathbf{D}(\mathbf{Z})}\mathrm{d}\mathbf{W}(t), \tag{2.41}$$

and its most probable deterministic dynamics:

$$\mathrm{d}\mathbf{z} = -\big(\mathbf{D}(\mathbf{z}) + \mathbf{Q}(\mathbf{z})\big)\nabla\varphi^s(\mathbf{z})\mathrm{d}t. \tag{2.42}$$

in the sense that the stationary distribution $p^s(\mathbf{z})$ of the stochastic system Eq. (2.29) and (2.41) corresponds to the Lyapunov function $\varphi(\mathbf{z}) = -\log(p^s(\mathbf{z}))$ of the most probable deterministic dynamics (2.42):

$$\frac{\mathrm{d}\varphi(\mathbf{z})}{\mathrm{d}t} = -\nabla\varphi^s(\mathbf{z})^{\mathrm{T}}\big(\mathbf{D}(\mathbf{z}) + \mathbf{Q}(\mathbf{z})\big)\nabla\varphi^s(\mathbf{z}) \leq 0. \tag{2.43}$$

In other words, the most probable dynamics directs the system towards states with higher stationary probability. We will return to this feature in Sec. 4.1 with a reaction diffusion equation model.

2.3.1 Corresponding Decomposition for Fokker-Planck equation

From Eq. (2.26), we can see that an alternative form of the Fokker-Planck equation (2.6) can be written, by plugging in the new form of $\mathbf{f}$, as: $\frac{\partial}{\partial t}p(\mathbf{z}, t) = \mathcal{L}\left[\frac{p(\mathbf{z}, t)}{\pi(\mathbf{z})}\right]$, where:

$$\mathcal{L}\left[\varphi(\mathbf{z})\right] = \nabla^{\mathrm{T}} \cdot \bigg(\big[\mathbf{D}(\mathbf{z}) + \mathbf{Q}(\mathbf{z})\big] \left[\left(\nabla\varphi(\mathbf{z})\right)\pi(\mathbf{z})\right] \bigg). \tag{2.44}$$

The operator $\mathcal{L}[\cdot]$ can be decomposed into a symmetric part $\mathcal{L}^S[\cdot]$, characterizing a reversible Markov process, and an anti-symmetric part $\mathcal{L}^A[\cdot]$, representing an irreversible process. The symmetric and skew-symmetric operators corresponds to two different kinds of dynamics.

The symmetric operator is determined solely by the diffusion matrix $\mathbf{D}$ since the skew-symmetric matrix $\mathbf{Q}$ cancels out:

$$\begin{aligned}
\mathcal{L}^S\left[\frac{p(\mathbf{z}|\mathbf{y};t)}{\pi(\mathbf{z})}\right] &= \frac{1}{2}\left(\mathcal{L}\left[\frac{p(\mathbf{z}|\mathbf{y};t)}{\pi(\mathbf{z})}\right] + \mathcal{L}^*\left[\frac{p(\mathbf{z}|\mathbf{y};t)}{\pi(\mathbf{z})}\right]\right) \\
&= \nabla^T \cdot \left(\mathbf{D}(\mathbf{z})\left[\nabla\left(\frac{p(\mathbf{z}|\mathbf{y};t)}{\pi(\mathbf{z})}\right)\pi(\mathbf{z})\right]\right) \\
&= \sum_{i,j=1}^{d} \frac{\partial^2}{\partial \mathbf{z}_i \partial \mathbf{z}_j}\Big(\mathbf{D}_{ij}(\mathbf{z})p(\mathbf{z}|\mathbf{y};t)\Big) \\
&+ \sum_{i=1}^{d} \frac{\partial}{\partial \mathbf{z}_i}\left(\left[\sum_j \mathbf{D}_{ij}(\mathbf{z})\frac{\partial H(\mathbf{z})}{\partial \mathbf{z}_j} - \Gamma_i^{\mathbf{D}}(\mathbf{z})\right]p(\mathbf{z}|\mathbf{y};t)\right).
\end{aligned} \tag{2.45}$$

Here, $\Gamma_i^{\mathbf{D}}(\mathbf{z}) = \sum_{j=1}^{d} \frac{\partial}{\partial \mathbf{z}_j}\mathbf{D}_{ij}(\mathbf{z})$ and $\mathcal{L}^*[\cdot]$ denotes the adjoint operator of $\mathcal{L}[\cdot]$. According to Itô's convention, the last two lines of (2.45) imply that $\frac{\partial}{\partial t}p(\mathbf{z}|\mathbf{y};t) = \mathcal{L}^S\left[\frac{p(\mathbf{z}|\mathbf{y};t)}{\pi(\mathbf{z})}\right]$ corresponds to reversible Brownian motion in a potential force field on a Riemannian manifold specified by the diffusion matrix $\mathbf{D}(\mathbf{z})$: $\mathrm{d}\mathbf{z} = \big[-\mathbf{D}(\mathbf{z})\nabla H(\mathbf{z}) + \Gamma^{\mathbf{D}}(\mathbf{z})\big]\mathrm{d}t + \sqrt{2\mathbf{D}(\mathbf{z})}\mathrm{d}\mathbf{W}(t)$. This is referred to as *Riemannian Langevin dynamics* [150]. When $\mathbf{D}(\mathbf{z})$ is positive definite, the reversible Markov dynamics have nice statistical regularity and will drive the system to converge to the stationary distribution.

The anti-symmetric operator is dictated solely by $\mathbf{Q}$, as here $\mathbf{D}$ cancels out:

$$\begin{aligned}\mathcal{L}^A\left[\frac{p(\mathbf{z}|\mathbf{y};t)}{\pi(\mathbf{z})}\right] =&\frac{1}{2}\left(\mathcal{L}\left[\frac{p(\mathbf{z}|\mathbf{y};t)}{\pi(\mathbf{z})}\right]-\mathcal{L}^*\left[\frac{p(\mathbf{z}|\mathbf{y};t)}{\pi(\mathbf{z})}\right]\right)\\ =&\nabla^T\cdot\left(\mathbf{Q}(\mathbf{z})\left[\nabla\left(\frac{p(\mathbf{z}|\mathbf{y};t)}{\pi(\mathbf{z})}\right)\pi(\mathbf{z})\right]\right)\\ =&\nabla^T\cdot\left(\Big[\mathbf{Q}(\mathbf{z})\nabla H(\mathbf{z})-\Gamma^{\mathbf{Q}}(\mathbf{z})\Big]p(\mathbf{z}|\mathbf{y};t)\right). \qquad (2.46)\end{aligned}$$

Here, $\Gamma_i^{\mathbf{Q}}(\mathbf{z})=\sum_{j=1}^d\frac{\partial}{\partial\mathbf{z}_j}\mathbf{Q}_{ij}(\mathbf{z})$. The last line of (2.46) demonstrates that $\frac{\partial}{\partial t}p(\mathbf{z}|\mathbf{y};t)=\mathcal{L}^A\left[\frac{p(\mathbf{z}|\mathbf{y};t)}{\pi(\mathbf{z})}\right]$ is a Liouville equation, which describes the density evolution of $p(\mathbf{z}|\mathbf{y};t)$ according to conserved, deterministic dynamics: $\mathrm{d}\mathbf{z}/\mathrm{d}t=-\mathbf{Q}(\mathbf{z})\nabla H(\mathbf{z})+\Gamma^{\mathbf{Q}}(\mathbf{z})$, with $\pi(\mathbf{z})$ its invariant measure.

2.4 Decomposition for Markov Jump Processes

We now turn our attention to the jump operator $\widehat{\mathcal{J}}[\cdot]$ of (2.5). As with the continuous dynamic operator $\widehat{\mathcal{L}}[\cdot]$, we consider an equivalent representation that separates the Markov jump process into symmetric and anti-symmetric components for more ready analysis of the properties of the process. The alternative representation $\mathcal{J}[\cdot]$ that we consider is defined in terms of two kernel functions S and A. A simple set of constraints on S and A ensures the equivalence between $\frac{\partial p(\mathbf{z}|\mathbf{y};t)}{\partial t}=\mathcal{J}\left[\frac{p(\mathbf{z}|\mathbf{y};t)}{\pi(\mathbf{z})}\right]$ and $\frac{\partial p(\mathbf{z}|\mathbf{y};t)}{\partial t}=\widehat{\mathcal{J}}\left[\frac{p(\mathbf{z}|\mathbf{y};t)}{\pi(\mathbf{z})}\right]$ where $\pi(\mathbf{z})$ is the stationary distribution of the jump process.

In particular, we consider

$$\mathcal{J}\left[\varphi(\mathbf{z})\right]=\int_{\mathbb{R}^d}\mathrm{d}\mathbf{x}\left[\big(S(\mathbf{x},\mathbf{z})+A(\mathbf{x},\mathbf{z})\big)\varphi(\mathbf{x})-S(\mathbf{x},\mathbf{z})\varphi(\mathbf{z})\right], \qquad (2.47)$$

where S is a symmetric kernel and A is an anti-symmetric kernel. Based on the form of (2.47), we simply have to satisfy the following constraints in order to ensure that $\pi(\mathbf{z})$ is the stationary distribution:

1. $\int_{\mathbb{R}^d}S(\mathbf{x},\mathbf{z})\pi^{-1}(\mathbf{x})\mathrm{d}\mathbf{x}$ and $\int_{\mathbb{R}^d}A(\mathbf{x},\mathbf{z})\pi^{-1}(\mathbf{x})\mathrm{d}\mathbf{x}$ exist

2. $S(\mathbf{x},\mathbf{z})+A(\mathbf{x},\mathbf{z})>0$

3. $\int_{\mathbb{R}^d}A(\mathbf{x},\mathbf{z})\mathrm{d}\mathbf{x}=0.$

Proof 2 *The reasoning is straightforward by introducing a symmetric kernel function* $S(\mathbf{x},\mathbf{z}) = S(\mathbf{z},\mathbf{x}) = \frac{1}{2}\Big(W(\mathbf{z}|\mathbf{x})\pi(\mathbf{x}) + W(\mathbf{x}|\mathbf{z})\pi(\mathbf{z})\Big)$*, and an anti-symmetric kernel* $A(\mathbf{x},\mathbf{z}) = -A(\mathbf{z},\mathbf{x}) = \frac{1}{2}\Big(W(\mathbf{z}|\mathbf{x})\pi(\mathbf{x}) - W(\mathbf{x}|\mathbf{z})\pi(\mathbf{z})\Big)$*, we arrive at a different form of* $\mathcal{J}[\cdot]$*:*

$$\mathcal{J}\left[\varphi(\mathbf{z})\right] = \int_{\mathbb{R}^d} d\mathbf{x} \left[S(\mathbf{x},\mathbf{z})\varphi(\mathbf{x}) - S(\mathbf{x},\mathbf{z})\varphi(\mathbf{z}) + A(\mathbf{x},\mathbf{z})\varphi(\mathbf{x})\right]. \tag{2.48}$$

Hence we can obtain the new formulation of the Markov jump process as:

$$\begin{aligned}\frac{\partial p(\mathbf{z}|\mathbf{y};t)}{\partial t} &= \mathcal{J}\left[\frac{p(\mathbf{z}|\mathbf{y};t)}{\pi(\mathbf{z})}\right] \\ &= \int_{\mathbb{R}^d} d\mathbf{x}\left[S(\mathbf{x},\mathbf{z})\frac{p(\mathbf{x}|\mathbf{y};t)}{\pi(\mathbf{x})} - S(\mathbf{x},\mathbf{z})\frac{p(\mathbf{z}|\mathbf{y};t)}{\pi(\mathbf{z})} + A(\mathbf{x},\mathbf{z})\frac{p(\mathbf{x}|\mathbf{y};t)}{\pi(\mathbf{x})}\right]. \end{aligned}\tag{2.49}$$

Plugging the stationary solution: $p(\mathbf{z}|\mathbf{y};t) = \pi(\mathbf{z})$ *into the above equation, we find the constraint on the anti-symmetric kernel:* $\int_{\mathbb{R}^d} A(\mathbf{x},\mathbf{z})d\mathbf{x} = 0$*. Since* $\frac{S(\mathbf{x},\mathbf{z}) + A(\mathbf{x},\mathbf{z})}{\pi(\mathbf{x})}$ *denotes a transition probability,* $S(\mathbf{x},\mathbf{z}) + A(\mathbf{x},\mathbf{z}) > 0$ *for any* $\mathbf{x}$ *and* $\mathbf{z}$*. We thereby notice that the requirement that* $\pi(\mathbf{z})$ *is a stationary distribution of the jump process is translated into simpler constraints:* $\int_{\mathbb{R}^d} S(\mathbf{x},\mathbf{z})\pi^{-1}(\mathbf{x})d\mathbf{x}$ *and* $\int_{\mathbb{R}^d} A(\mathbf{x},\mathbf{z})\pi^{-1}(\mathbf{x})d\mathbf{x}$ *exists, with* $S(\mathbf{x},\mathbf{z}) + A(\mathbf{x},\mathbf{z}) > 0$*, and* $\int_{\mathbb{R}^d} A(\mathbf{x},\mathbf{z})d\mathbf{x} = 0$*.*

Following (2.10), the transition probability implied by the operator of (2.47) assuming a Δt-discretization is given by:

$$p(\mathbf{z}|\mathbf{y};\Delta t) = \frac{\Delta t}{\pi(\mathbf{y})}\big(S(\mathbf{y},\mathbf{z}) + A(\mathbf{y},\mathbf{z})\big) + \left[1 - \frac{\Delta t}{\pi(\mathbf{y})}\int_{\mathbb{R}^d} S(\mathbf{y},\mathbf{x})\mathrm{d}\mathbf{x}\right]\delta(\mathbf{z}-\mathbf{y}). \tag{2.50}$$

2.4.1 Reversible and Irreversible Dynamics of $\mathcal{J}[\cdot]$

Similar to operator $\mathcal{L}[\cdot]$, operator $\mathcal{J}[\cdot]$ can also be decomposed into a symmetric (reversible) part $\mathcal{J}^S[\cdot]$ and anti-symmetric (irreversible) part $\mathcal{J}^A[\cdot]$:

$$\mathcal{J}^S\left[\varphi(\mathbf{z})\right] = \frac{1}{2}\left(\mathcal{J}\left[\varphi(\mathbf{z})\right] + \mathcal{J}^*\left[\varphi(\mathbf{z})\right]\right) = \int_{\mathbb{R}^d} \mathrm{d}\mathbf{x}\left[S(\mathbf{x},\mathbf{z})\varphi(\mathbf{x}) - S(\mathbf{x},\mathbf{z})\varphi(\mathbf{z})\right]; \tag{2.51}$$

$$\mathcal{J}^A\left[\varphi(\mathbf{z})\right] = \frac{1}{2}\left(\mathcal{J}\left[\varphi(\mathbf{z})\right] - \mathcal{J}^*\left[\varphi(\mathbf{z})\right]\right) = \int_{\mathbb{R}^d} \mathrm{d}\mathbf{x}\left[A(\mathbf{x},\mathbf{z})\varphi(\mathbf{x})\right]. \tag{2.52}$$

Here, $\mathcal{J}^*[\cdot]$ is the adjoint operator of $\mathcal{J}[\cdot]$. We see that A fully determines the irreversible dynamics whereas S defines the reversible part. We can further derive from (2.50) that A is the difference between the probability of a forward path and the backward path:

$$A(\mathbf{x},\mathbf{z}) = \frac{1}{2\Delta t}\big(\pi(\mathbf{y})p(\mathbf{z}|\mathbf{y};\Delta t) - \pi(\mathbf{z})p(\mathbf{y}|\mathbf{z};\Delta t)\big). \tag{2.53}$$

Chapter 3

GENERALIZING THERMODYNAMICS IN STOCHASTIC PROCESSES

Traditionally, thermodynamics has been defined for high dimensional Hamiltonian systems to capture the global behaviors of them. It is one of the goals of this thesis to extend it to general stochastic processes.

In this chapter, we first examine the traditional theory of Boltzmann and Gibbs with a direct generalization of it to systems with conservation laws to see that their theory is contingent upon the systems possessing a conserved quantity. In Sec. 3.1, we will see that thermodynamics can be naturally defined for the Lotka-Volterra (LV) equations and provide new vocabularies to describe ecological systems from a global point of view. As a non-Hamiltonian system, the LV equations bring in richer structure to the formulation of the thermodynamic theory. Invariant measure of the system plays a central role in the new conservation laws. More importantly, the ecological systems possess a crucial feature, absent in the classical mechanics, a natural stochastic population dynamic formulation of which the deterministic equation (e.g., the LV equation studied) is the infinite population limit. Studies of the stochastic dynamics with finite populations show the LV equation as the robust, fast cyclic underlying behavior. Hence the thermodynamic theory can be directly introduced into stochastic systems.

We then discuss this thermodynamic theory with general irreversible stochastic systems. When a natural separation of time scale is not present, decomposition from Sec. 2.3 can be used. In Sec. 3.2, we revisit the Ornstein-Uhlenbeck (OU) process as the fundamental mathematical description of linear irreversible phenomena, with fluctuations, near equilibrium. We formulate the thermodynamic theory again through identifying the underlying circulating dynamics in a stationary process as the natural generalization of classical conservative mechanics. A bridge between a family of OU processes with equilibrium fluctuations and thermodynamics is thus established through the celebrated Helmholtz theorem. The Helmholtz theorem provides the emergent macroscopic “equation of state” of the entire system, which exhibits a universal ideal thermodynamic behavior, parallel to that of ideal gas. Fluctuating macroscopic quantities are studied from the stochastic thermodynamic point of view and a non-equilibrium work relation is obtained in the macroscopic picture, which may facilitate experimental study and application of the equalities due to Jarzynski, Crooks, and Hatano and Sasa.

3.1 Thermodynamics from Conservation Laws

In the modern days' language, Boltzmann and Gibbs tried to characterize the global behaviors of a Hamiltonian dynamical system as a function of parameters and initial conditions. The central idea is to use the fact that given a certain initial condition $\mathbf{x}_0$ and parameter α, all the states of the system is aligned on a single level set of $\phi(\mathbf{x}, \alpha) = E$. Then given a certain upper bound of energy level, all the states are confined inside a compact set $\mathcal{S}(E, \alpha) = \{\mathbf{x}|\phi(\mathbf{x}, \alpha) \leq E\}$, assuming that function $\phi(\mathbf{x}, \alpha)$ is proper in the space of $\mathbf{x}$. A characterization of the geometry of the set $\mathcal{S}$ reflects the overall behavior of an entire trajectory (or class of trajectories) as a non-constant but steady state. In classical mechanics, the "Boltzmann-Gibbs entropy" is defined as the log of the volume of $\mathcal{S}$ to quantify its geometry:

$$\sigma_B(E, \alpha) = \ln \left(\int_{\phi(\mathbf{x},\alpha)\leq E} \mathrm{d}\mu(\mathbf{x}) \right),$$

where measure $\mathrm{d}\mu(\mathbf{x})$ can be taken as either Lebesgue measure according to Boltzmann and Gibbs' original construction, or the system's invariant measure to quantify the effective phase space volume that the trajectories of the system would explore. With $\sigma_B(E, \alpha)$ an increasing function of E, let's further assume that the implicit function theorem applies and write:

$$E = E(\sigma_B, \alpha).$$

In differential form,

$$\begin{aligned} \mathrm{d}E &= \theta(E, \alpha)\mathrm{d}\sigma_B - F_\alpha(E, \alpha)\mathrm{d}\alpha \\ &= \left(\frac{\partial \sigma_B}{\partial E}\right)^{-1} \mathrm{d}\sigma - \left(\frac{\partial \sigma_B}{\partial \alpha}\right) \left(\frac{\partial \sigma_B}{\partial E}\right)^{-1} \mathrm{d}\alpha. \end{aligned} \tag{3.1}$$

The above equation is called the Helmholtz theorem. The two conjugate variables, θ and F_α, correspond to the macroscopic quantities in classical thermodynamics as temperature and force. The force here should be understood as Onsager's thermodynamic force: corresponding to a spatial displacement is a mechanical force; to a change in number of particles is Gibbs' chemical potential; to a variation in a parameter through a Maxwell demon then is an informatic force [168, 106].

The thermodynamic conjugate variable of α, the α-force:

$$F_\alpha = \theta \cdot \left(\frac{\partial \sigma_B(E, \alpha)}{\partial \alpha}\right)_E. \tag{3.2}$$

A mathematical relation between α, F_α, and θ is called an *equation of state* in classical thermodynamics. It is important to note that the above arguments holds for any system with a conservative quantity.

In this section, we take the celebrated Lotka-Volterra equation as a motivating example and analyze its thermodynamics. It is particularly suitable in that it is a system with conserved dynamics, and at the same time not a Hamiltonian system. Furthermore, studies of the stochastic dynamics with finite populations show the Lotka-Volterra equation as the robust, fast cyclic underlying behavior, which paves the way for our study of the thermodynamics in general stochastic dynamics.

3.1.1 Lotka-Volterra equation and population dynamics

Among ecological models, the Lotka-Volterra (LV) equation for predator-prey system has played an important pedagogical role [109, 83, 100], even though it is certainly not a realistic model for any engineering applications. We choose this population system in the present work because its mathematics tractability, and its stochastic counterpart in terms of a birth and death process [3, 59]. It can be rigorously shown that a smooth solution to the Lotka-Volterra differential equation is the law of large numbers for the stochastic process [86]. In biochemistry, the birth-death process for discrete, stochastic reactions corresponding to the mass-action kinetics has been called a Delbrück-Gillespie process [133].

In its non-dimensionalized form, the Lotka-Volterra equation reads [109]:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = x(1-y) = f(x,y), \;\; \frac{\mathrm{d}y}{\mathrm{d}t} = \alpha y(x-1) = g(x,y;\alpha), \tag{3.3}$$

in which $x(t)$ and $y(t)$ represent the populations of a prey and its predator, each normalized with respect to its time-averaged mean populations. The xy term in $f(x,y)$ stands for the rate of consumption of the prey by the predator, and the αyx term in $g(x,y;\alpha)$ stands for the rate of prey-dependent predator growth.

It is easy to check that the solutions to (3.3) in phase space are level curves of a scalar function [109]

$$H(x,y) = \alpha x + y - \ln\big(x^\alpha y\big). \tag{3.4}$$

We shall use $\Gamma_{H=h}$ to denote the solution curve $H(x,y) = h$, and $\mathfrak{D}_h(\alpha)$ to denote the domain encircled by the $\Gamma_{H=h}$. Fig. 4.1 shows the contours of $H(x,y)$ with $\alpha = 1$ and $H(x,y) = 2.61$ with different α's.

Let τ be the period of the cyclic dynamics. Then it is easy to show that [109]

$$\frac{1}{\tau}\int_0^\tau x(t)\mathrm{d}t = \frac{1}{\tau}\int_0^\tau y(t)\mathrm{d}t = 1. \tag{3.5}$$

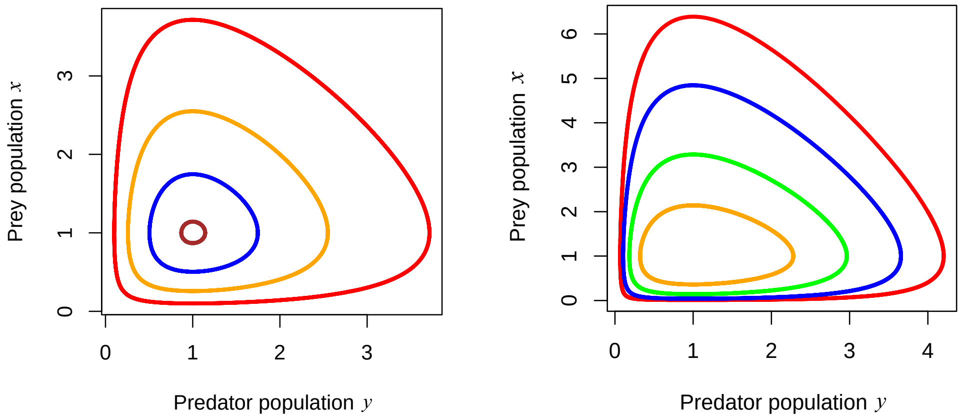


Figure 3.1: Left panel: with $\alpha = 1$ and $H(x,y) = 3.40, 2.61, 2.19$, and 2.01. Right panel: with $\alpha = 0.5, 0.6, 0.8$, and 1.2, from outside inward, all with $H(x,y) = 2.61$. We see that the larger the α, the smaller the temporal variations in the prey population, relative to that of predator.

Furthermore (see Appendix A),

$$\frac{1}{\tau}\int_0^{\tau} \big(x(t)-1\big)^2 \mathrm{d}t = \frac{\hat{\mathcal{A}}}{\alpha\tau}, \tag{3.6}$$

$$\frac{1}{\tau}\int_0^{\tau} \big(y(t)-1\big)^2 \mathrm{d}t = \frac{\alpha\hat{\mathcal{A}}}{\tau}, \tag{3.7}$$

in which $\hat{\mathcal{A}}$ is the area of $\mathfrak{D}_h(\alpha)$, encircled by $\Gamma_{H=h}$, using Lebesgue measure in the xy-plane. The appropriate measure for computing the area will be further discussed in Sec. 3.1.2. The parameter α represents the relative temporal variations, or dynamic ranges, in the two populations: the larger the α, the greater the temporal variations and range in the predator population, and the smaller in the prey population.

3.1.2 The Helmholtz theorem

Eq. (3.3) is not a Hamiltonian system, nor is it divergence-free

$$\frac{\partial f(x,y)}{\partial x} + \frac{\partial g(x,y;\alpha)}{\partial y} \neq 0.$$

It can be expressed, however, as

$$\begin{pmatrix} \mathrm{d}x/\mathrm{d}t \\ \mathrm{d}y/\mathrm{d}t \end{pmatrix} = \begin{pmatrix} 0 & -G(x,y) \\ G(x,y) & 0 \end{pmatrix} \nabla\phi(x,y) non-df. \tag{3.8}$$

with a scalar factor $G(x,y) = xy$. One can in fact understand this scalar factor as a "local change of measure", or time $\mathrm{d}\hat{t} \equiv G\big(x(t),y(t)\big)\mathrm{d}t$ [134]:

$$x(t) = \hat{x}\big(\hat{t}(t)\big),\ \ y(t) = \hat{y}\big(\hat{t}(t)\big), \tag{3.9}$$

for

$$\hat{t}(t) = \hat{t}_0 + \int_{t_0}^{t} G\big(x(s),y(s)\big)\mathrm{d}s,$$

where $(\hat{x},\hat{y})$ satisfies the corresponding Hamiltonian system. In Sec. 3.1.3 and 3.1.4 below, we shall show that $G^{-1}(x,y)$ is an invariant density of the Liouville equation for the deterministic dynamics (3.3), and more importantly the invariant density of the Fokker-Planck equation for the corresponding stochastic dynamics. As will be demonstrated in Sec. 3.1.2 and 3.1.2, statistical average of quantities according to the invariant measure $G^{-1}(x,y)\mathrm{d}x\mathrm{d}y$ can be calculated through time average of those quantities along the system's instantaneous dynamics. Knowledge about the system's long term distribution is not needed during the calculation. These facts make the $G^{-1}(x,y)$ the natural measure for computing area $\mathcal{A}$.

Any function of $H(x,y)$, $\rho(H)$ is conserved under the dynamics, as is guaranteed by the orthogonality between the vector field of (3.3) and gradient $\nabla\rho$ [**?**]:

$$\begin{aligned} \frac{\mathrm{d}\rho\big(H(x,y)\big)}{\mathrm{d}t} &= f(x,y)\frac{\partial}{\partial x}\rho\big(H(x,y)\big) + g(x,y;\alpha)\frac{\partial}{\partial y}\rho\big(H(x,y)\big) \\ &= \rho'(H)\left(x(1-y)\alpha\left(1-\frac{1}{x}\right) + \alpha y(x-1)\left(1-\frac{1}{y}\right)\right) \\ &= 0. \end{aligned} \tag{3.10}$$

This is analogous to the "conservation law" observed in Hamiltonian systems.

Extending the conservation law

The nonlinear dynamics in (3.3), therefore, introduces a "conservative relation" between the populations of predator and prey according to (3.4). If we call the value $H(x,y)$ an "energy", then the phase portrait in the left panel of Fig. 4.1 suggests that the entire phase space of the dynamical system is organized according to the value of H. The deep insight contained in the work of Helmholtz and Boltzmann [49, 25] is that such an energy-based organization can be further extended for different values of α: Therefore, the energy-based organization is no longer limited to a *single* orbit, nor a *single* dynamical system; but rather for the entire class of parametric dynamical systems. In the classical physics of Newtonian mechanical energy conservation, this yields the mechanical basis of the Fundamental Thermodynamic Relation as a form of the First Law, which extends the notion of energy conservation far beyond mechanical systems [126, 122].

More specifically, we see that the area $\mathcal{A}$ in Fig. 4.1, or in fact any geometric quantification of a closed orbit, is completely determined by the parameter α and initial energy value $h = H\big(x(0), y(0), \alpha\big)$. Therefore, there must exist a bivariate function $\mathcal{A} = \mathcal{A}(h, \alpha)$, Assuming the implicit function theorem applies, then one has

$$h = h(\mathcal{A}, \alpha). \tag{3.11}$$

Note that in terms of the Eq. (3.11), a "state" of the ecological system is not a single point (x, y) which is continuously varying with time; rather it reflects the geometry of an entire orbit. Then Eq. (3.11) implies that any such ecological state has an "h-energy", *if* one recognizes a geometric, state variable $\mathcal{A}$.

Eq. (3.11) can be written in a differential form

$$\mathrm{d}h = \left(\frac{\partial h}{\partial \mathcal{A}}\right)_{\alpha} \mathrm{d}\mathcal{A} + \left(\frac{\partial h}{\partial \alpha}\right)_{\mathcal{A}} \mathrm{d}\alpha, \tag{3.12}$$

in which one first introduces the h-energy for an ecological system with fixed α via the factor $(\partial h/\partial \mathcal{A})$. Then, holding $\mathcal{A}$ constant, one introduces an "α-force" corresponding to the parameter α. In classical thermodynamics, the latter is known as an "adiabatic" process.

The Helmholtz theorem expresses the two partial derivatives in (3.12) in terms of the dynamics in Eq. (3.3).

Projected invariant measure

For canonical Hamiltonian systems, Lebesgue measure is an invariant measure in the whole phase space. On the level set $\Gamma_{H=h}$, the projection of the Lebesgue measure, called the Liouville measure, also defines an invariant measure on the sub-manifold. If the dynamics on the invariant sub-manifold $\Gamma_{H=h}$ is ergodic, the average with respect to the Liouville measure is equal to the time average along the trajectory starting from any initial condition (x_0, y_0) satisfying $H(x_0, y_0) = h$.

As we shall show below, the invariant measure for the LV system (3.3) in the whole phase space is $\mathrm{d}\mathcal{A} = G^{-1}(x,y)\mathrm{d}x\mathrm{d}y$. Projection of this invariant measure onto the level set $\Gamma_{H=h}$ can be found by changing (x,y) to intrinsic coordinates (h,ℓ):

$$\mathrm{d}\mathcal{A} = G^{-1}(x,y)\;\mathrm{d}x\mathrm{d}y = G^{-1}(x,y)\;\big(\mathrm{d}x, \mathrm{d}y\big)^T \cdot \mathbf{n}\,\mathrm{d}\ell, \tag{3.13}$$

where

$$\mathbf{n} = \left(\frac{\partial H(x,y)/\partial x}{||\nabla H(x,y)||}, \frac{\partial H(x,y)/\partial y}{||\nabla H(x,y)||}\right)^T_{(x,y)\in\Gamma_{H=h}} \tag{3.14}$$

is the unit normal vector of the the level set $\Gamma_{H=h}$; and $\mathrm{d}\ell = \sqrt{\mathrm{d}x^2 + \mathrm{d}y^2}$. Noting that:

$$\mathrm{d}h \;\;=\;\; \frac{\partial H(x,y)}{\partial x}\mathrm{d}x + \frac{\partial H(x,y)}{\partial y}\mathrm{d}y, \tag{3.15}$$

we have

$$\big(\mathrm{d}x, \mathrm{d}y\big)^T \cdot \mathbf{n} = \frac{\mathrm{d}h}{||\nabla H(x,y)||}. \tag{3.16}$$

That is:

$$\mathrm{d}\mathcal{A} = \mathrm{d}\mu\;\mathrm{d}h, \tag{3.17}$$

where

$$\mathrm{d}\mu = \frac{G^{-1}(x,y)}{||\nabla H(x,y)||}\mathrm{d}\ell \tag{3.18}$$

is the projected invariant measure of the Lotka-Volterra system on the level set $\Gamma_{H=h}$.

It is worth noting that $\mathrm{d}\mu = \mathrm{d}t$ on the level set $\Gamma_{H=h}$. Since dynamics on $\Gamma_{H=h}$ is ergodic, the average of any function $\psi(x,y)$ under the projected invariant measure on $\Gamma_{H=h}$ is equal to its time average over a period:

$$\begin{aligned}\langle\psi\rangle^{\Gamma_{H=h}} &\triangleq \frac{\displaystyle\oint_{\Gamma_{H=h}} \psi\big(x,y\big)\mathrm{d}\mu}{\displaystyle\oint_{\Gamma_{H=h}} \mathrm{d}\mu} \\ &= \frac{1}{\tau}\int_0^\tau \psi\big(x(t), y(t)\big)\mathrm{d}t \triangleq \langle\psi\rangle^t \end{aligned} \tag{3.19}$$

Functional relation between $\ln \boldsymbol{\mathcal{A}}$*,* $\boldsymbol{\alpha}$*, and* $\boldsymbol{h}$

Under the invariant measure $G^{-1}(x,y)$, the area $\mathcal{A}$ encircled by the level curve $\Gamma_{H=h}$ is:

$$\begin{aligned}\mathcal{A}_{\mathfrak{D}_h(\alpha)} &= \iint_{\mathfrak{D}_h(\alpha)} G^{-1}(x,y)\mathrm{d}x\mathrm{d}y \\ &= \iint_{\mathfrak{D}_h(\alpha)} \mathrm{d}\ln x\,\mathrm{d}\ln y \end{aligned} \tag{3.20}$$

Using Green's theorem the area $\mathcal{A}_{\mathfrak{D}_h(\alpha)}$ can be simplified as

$$\begin{aligned}\mathcal{A}_{\mathfrak{D}_h(\alpha)} &= \int_0^{\tau(h,\alpha)} \ln y \left(\frac{\partial H}{\partial \ln y}\right) \mathrm{d}t \\ &= \int_0^{\tau(h,\alpha)} \ln x \left(\frac{\partial H}{\partial \ln x}\right) \mathrm{d}t, \end{aligned} \tag{3.21}$$

where $\tau(h,\alpha)$ is the time period for the cyclic motion. Furthermore,

$$\begin{aligned}\frac{\partial \mathcal{A}_{\mathfrak{D}_h(\alpha)}}{\partial h} &= \frac{\partial}{\partial h}\iint_{\mathfrak{D}_h(\alpha)} G^{-1}(x,y)\,\mathrm{d}x\mathrm{d}y \\ &= \int_0^{\tau(h,\alpha)} \mathrm{d}t = \tau(h,\alpha). \end{aligned} \tag{3.22}$$

That is

$$\left(\frac{\partial \ln \mathcal{A}_{\mathfrak{D}_h(\alpha)}}{\partial h}\right)^{-1} = \left\langle \ln x \left(\frac{\partial H}{\partial \ln x}\right)\right\rangle^t = \left\langle \ln y \left(\frac{\partial H}{\partial \ln y}\right)\right\rangle^t, \tag{3.23}$$

in which $\langle\cdots\rangle^t$ is the time average, or phase space average according to the invariant measure. We can also find the derivative of the area $\mathcal{A}_{\mathfrak{D}_h(\alpha)}$ encircled by the level curve $\Gamma_{H=h}$ with respect to the parameter of the system α as:

$$\begin{aligned}\frac{\partial \mathcal{A}_{\mathfrak{D}_h(\alpha)}}{\partial \alpha} &= \frac{\partial}{\partial \alpha}\iint_{\mathfrak{D}_h(\alpha)} G^{-1}(x,y)\,\mathrm{d}x\mathrm{d}y \\ &= -\int_0^{\tau(h,\alpha)} \big(x(t) - \ln x(t)\big)\,\mathrm{d}t. \end{aligned} \tag{3.24}$$

In this setting, the Helmholtz theorem reads

$$\mathrm{d}h = \frac{\mathrm{d}\mathcal{A} - \left(\dfrac{\partial \mathcal{A}}{\partial \alpha}\right)_h \mathrm{d}\alpha}{\left(\dfrac{\partial \mathcal{A}}{\partial h}\right)_\alpha} = \theta(h,\alpha)\,\mathrm{d}\ln\mathcal{A} - F_\alpha(h,\alpha)\mathrm{d}\alpha, \tag{3.25}$$

in which

$$\theta(h,\alpha) = \mathcal{A}_{\mathfrak{D}_h(\alpha)}\left(\frac{\partial\mathcal{A}}{\partial h}\right)_{\alpha}^{-1} = \left\langle \ln x\left(\frac{\partial H}{\partial \ln x}\right)\right\rangle^{t} = \left\langle \ln y\left(\frac{\partial H}{\partial \ln y}\right)\right\rangle^{t}. \tag{3.26}$$

The factor $\theta(h,\alpha)$ here is the mean $\ln x(\partial H/\partial \ln x)$, or $\ln y(\partial H/\partial \ln y)$, precisely like the mean kinetic energy as the notion of temperature in classical physics, and the virial theorem. The α-force is then defined as

$$F_{\alpha}(h,\alpha) = \left(\frac{\partial\mathcal{A}}{\partial \alpha}\right)_{h}\left(\frac{\partial\mathcal{A}}{\partial h}\right)_{\alpha}^{-1} = -\left\langle\frac{\partial H(x,y,\alpha)}{\partial\alpha}\right\rangle^{t}. \tag{3.27}$$

It is important to note that the definition of F_α given in the right-hand-side of (3.27) is completely independent of the notion of $\mathcal{A}$, even though the relation (3.25) explicitly involves the latter. $F_\alpha(h,\alpha)$ is a function of both h and α, however. Therefore, the value of α-work $F_\alpha(h,\alpha)\mathrm{d}\alpha$ depends on how h is constrained: There are iso-h processes, iso-θ processes, etc. [136]

Equation of state

The notion of an equation of state first appeared in classic thermodynamics [126, 122]. From a modern dynamical systems standpoint, a fixed point as a function usually is continuously dependent upon the parameters in a mathematical model, except at bifurcation points. Let $(x_1^*, x_2^*, \cdots, x_n^*)$ be a globally asymptotically attractive fixed point, and α be a parameter, then the function $x_1^*(\alpha)$ constitutes an *equation of state* for the long-time "equilibrium" behavior of the dynamical system.

If a system has a globally asymptotically attractive limit set that is not a simple fixed point, then every geometric characteristic of the invariant manifold, say $\mathfrak{g}^*$, will be a function of α. In this case, $\mathfrak{g}^*(\alpha)$ could well be considered as an equation of state. An "equilibrium state" in this case is the entire invariant manifold.

The situation for a conservative dynamical system with center manifolds is quite different. In this case, the long-time behavior of the dynamical system, the foremost, is dependent upon its initial data. An equation of state therefore is a functional relation among (i) geometric characteristics of a center manifold $\mathfrak{g}^*$, (ii) parameter α, and (iii) a new quantity, or quantities, that identifies a specific center manifold, h. This is the fundamental insight of the Helmholtz theorem.

In ecological terms, area under the invariant measure: $\mathcal{A}$, gives a sense of total variation in both the predator's and the prey's populations. Therefore, $\ln\mathcal{A}$ measures population range of both populations as a whole. The parameter α, on the other hand, is the proportion of predators' over preys' population ranges of time variations:

$$\alpha^2 = \frac{\int_0^{\tau(h,\alpha)} \big(y(t)-1\big)^2 \mathrm{d}t}{\int_0^{\tau(h,\alpha)} \big(x(t)-1\big)^2 \mathrm{d}t} = \frac{\langle (y-1)^2 \rangle^t}{\langle (x-1)^2 \rangle^t}. \tag{3.28}$$

The new quantity θ can be viewed as a measure of the mean ecological "activeness":

$$\theta = \langle \alpha(x-1)\ln x \rangle^t = \langle (y-1)\ln y \rangle^t. \tag{3.29}$$

It is the mean of "distance" from the prey's and predator's populations x and y, to the fixed populations in equilibrium $(1,1)$. For population dynamic variable u, Eq. 3.29 suggests a norm $\|u\| \equiv u\ln(u+1)$. Then, $\theta = \langle \alpha\|x-1\| \rangle^t = \langle \|y-1\| \rangle^t$; and an averaged norm of per capita growth rates in the two species:

$$\theta = \left\langle \alpha \left\| \frac{1}{\alpha}\frac{\mathrm{d}\ln y}{\mathrm{d}t} \right\| \right\rangle^t = \left\langle \left\| -\frac{\mathrm{d}\ln x}{\mathrm{d}t} \right\| \right\rangle^t. \tag{3.30}$$

And finally,

$$F_\alpha = -\left\langle \frac{\partial H(x,y,\alpha)}{\partial \alpha} \right\rangle^t = -\langle x - \ln x \rangle^t \tag{3.31}$$

is the "ecological force" one needs to counteract in order to change α. In other words, when $|F_\alpha|$ is greater, more h-energy change is needed to vary α. It is also worth noting that $|F_\alpha|$ is positively related to the prey's average population range. In fact we can define another "distance" of the prey's population x to 1 as: $\|u\|_F = u - \ln(u+1)$, then $F_\alpha = -\big\langle \|x-1\|_F \big\rangle^t - 1$. Note that for very small u: $\|u\| \approx u^2 \approx 2\|u\|_F$

Fig. 4.2 shows various forms of "the equation of state", e.g., relationships among the triplets $\big(\alpha, |F_\alpha| = -F_\alpha, h\big)$, $\big(|F_\alpha|, \theta, h\big)$, and $\big(\alpha, \theta, h\big)$ in the first row; among the triplet $\big(\alpha, |F_\alpha|, \theta\big)$ in the second row; and among the triplet $\big(\mathcal{A}, \theta, \alpha\big)$ in the third row. The second row shows that the relation among $\big(\alpha, |F_\alpha|, \theta\big)$ is just like that among $\big(V, P, T\big)$ in ideal gas model: Mean ecological activeness θ increases nearly linearly with the ecological force $|F_\alpha|$ for constant α, and with the proportion α of the predator's population range over the prey's, for constant $|F_\alpha|$; Ecological force $|F_\alpha|$ and the proportion α of population ranges are inversely related under constant mean activeness θ. And when $\theta = 0$, $\alpha(F_\alpha + 1) = 0$. Other features can be observed by looking into the details of each column.

The first column of Fig. 4.2 demonstrates that as the proportion α of population ranges increases, the ecological force $|F_\alpha|$ is alleviated (for given h-energy or ecological activeness θ). This is due to the positive relationship between the ecological force $|F_\alpha|$ and the prey's population range (as shown in Eq. 3.31). Since α is the proportion of the predator's

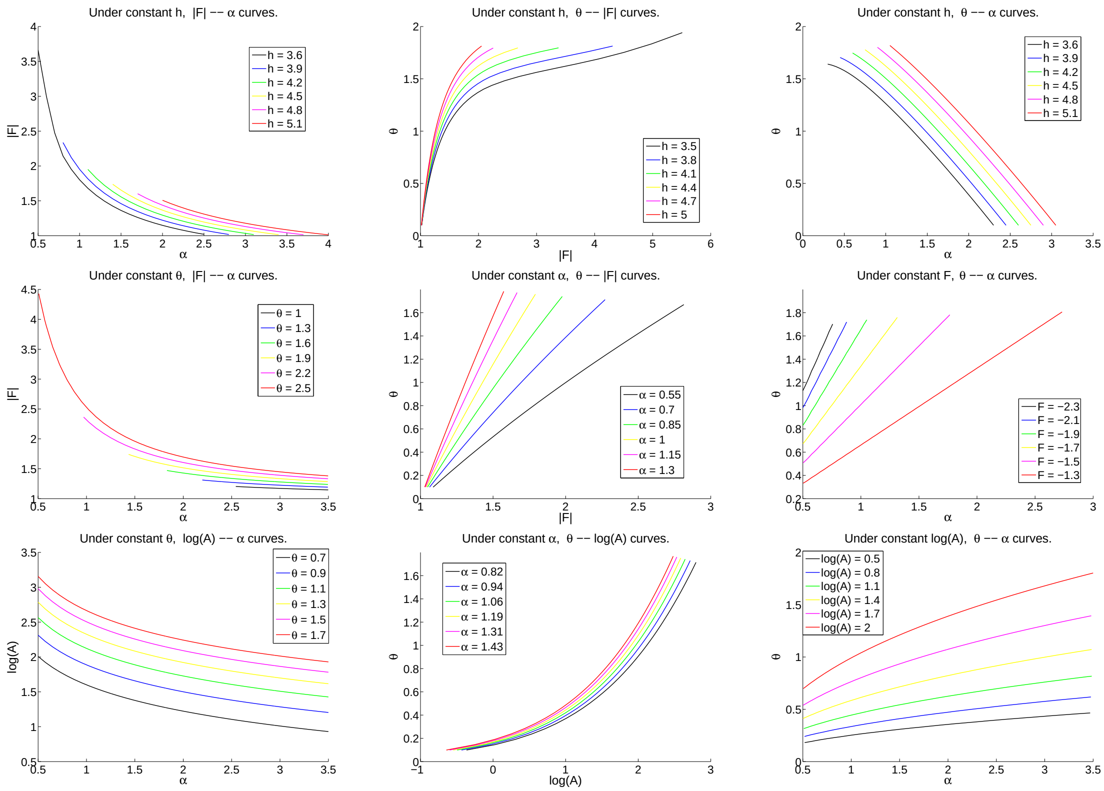


Figure 3.2: Various functional relationships, e.g., "the equation of state", among $\big(|F|, \alpha, h\big)$, $\big(|F|, \theta, h\big)$, and $\big(\theta, \alpha, h\big)$ in the top row, different views among $\big(|F|, \theta, \alpha\big)$ in the second row, and among $\big(\ln(\mathcal{A}), \theta, \alpha\big)$ in the third row.

population range over the prey's, $|F_\alpha|$ and α would be inversely related when any resource, h-energy, or activity, θ, remains constant. This fact means that on an iso-h or iso-θ curve, when the proportion α is large, relatively less h-energy change is needed to reduce it. The first column also demonstrates an inverse relationship between α and the total population range $\ln \mathcal{A}$ for any given θ, which reflects the fact that as the proportion of the predator's population range over the prey's increase, the total population range of the two species would actually decrease.

The second column of Fig. 4.2 demonstrates that: the ecological force $|F_\alpha|$ and the total population range $\ln \mathcal{A}$ increases with the mean activeness θ (with given h-energy or α). This observation means that it would also take more h-energy to change the proportion

α of the predator's population range over the prey's, if mean ecological activeness rises, and that more population range would be explored with more ecological activeness θ.

The third column about the relation between θ and α is interesting: Under constant h-energy, as the proportion α of population ranges increases, the ecological activeness θ decreases, in accordance with the drop in the total population range $\ln \mathcal{A}$ as shown in Fig 4.1. But when the total population range $\ln \mathcal{A}$ or the ecological force F_α is to remain constant, ecological activeness θ actually increases with α. This means that under constant resource (h-energy), the proportion α of the predator's population range over the prey's restricts mean ecological activeness. But if we fix the ecological force or total population range (supplying more h-energy), an increase in predator's population range over prey's can increase ecological activeness.

3.1.3 Liouville description in phase space

Nonlinear dynamics described by Eq. (3.3) has a linear, first-order partial differential equation (PDE) representation

$$\frac{\partial u(x,y,t)}{\partial t} = -\frac{\partial}{\partial x}\Big(f(x,y)u(x,y,t)\Big) - \frac{\partial}{\partial y}\Big(g(x,y;\alpha)u(x,y,t)\Big). \tag{3.32}$$

A solution to (3.32) can be obtained via the method of characteristics, exactly via (3.3). Eq. (3.32) sometime is called the Liouville equation for the ordinary differential equations (3.3). It also has an adjoint:

$$\frac{\partial v(x,y,t)}{\partial t} = f(x,y)\frac{\partial v(x,y,t)}{\partial x} + g(x,y;\alpha)\frac{\partial v(x,y,t)}{\partial y}. \tag{3.33}$$

Note that while the orthogonality in Eq. (3.10) indicates that $\rho\big(H(x,y)\big)$ is a stationary solution to Eq. (3.33), it is not a stationary invariant density to (3.32).

This is due to the fact that vector field (f,g) is not divergence free, but rather as in (**??**) the scalar factor $G(x,y) = xy$. Then it is easy to verify that $G^{-1}(x,y)\rho(H(x,y))$ is a stationary solution to (3.32):

$$\frac{\partial}{\partial x}\left(f(x,y)\,\frac{\rho(H(x,y))}{G(x,y)}\right) + \frac{\partial}{\partial y}\left(g(x,y)\,\frac{\rho(H(x,y))}{G(x,y)}\right) = 0. \tag{3.34}$$

Entropy dynamics in phase space

It is widely known that a volume-preserving, divergence-free conservative dynamics has a conserved entropy $S[u(x,t)] = -\int_{\mathbb{R}} u(x,t)\ln u(x,t)\mathrm{d}x$ [7]. For conservative system like

(3.3) which contains the scalar factor $G(x, y)$, the Shannon entropy should be replaced by the relative entropy with respect to $G^{-1}(x, y)$ (see Appendix B for detailed calculation):

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^2} u(x,y,t)\ln\left(\frac{u(x,y,t)}{G^{-1}(x,y)}\right)\mathrm{d}x\mathrm{d}y = 0. \tag{3.35}$$

Such systems are called *canonical conservative* with respect to $G^{-1}(x, y)$ in [134]. In classical statistical physics, the term $\int_{\mathbb{R}} u\ln\left(u/G^{-1}\right)\mathrm{d}x$ is called free energy [188]; in information theory, Kullback-Leibler divergence.

We can in fact show a stronger result, with arbitrary differentiable $\Psi(\cdot)$ and $\rho(\cdot)$ over an arbitrary domain $\mathfrak{D}$ (see Appendix B):

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathfrak{D}} u(x,y,t)\Psi\left(\frac{u(x,y,t)}{G^{-1}(x,y)\rho(H(x,y))}\right)\mathrm{d}x\mathrm{d}y\\ =\ &\int_{\partial\mathfrak{D}}\left\{u(x,y,t)\Psi\left(\frac{u(x,y,t)}{G^{-1}(x,y)\rho(H)}\right)\big(f,g\big)\right\}\times(\mathrm{d}x,\mathrm{d}y).\end{aligned} \tag{3.36}$$

Therefore, if $\mathfrak{D} = \mathfrak{D}_h$, then $\partial\mathfrak{D} = \Gamma_{H=h}$, and the integral on the right-hand-side of (3.36) is always zero. In other words, in conservative dynamics like (3.3), it is the support $\mathfrak{D} \subset \mathbb{R}^2$ on which $u(x, y, t)$ is observed that determines whether a system is invariant; not the initial data $u(x, y, 0)$ [**?**].

Relation between $\mathcal{A}$, Shannon entropy, and relative entropy

Since a "state" is defined as an entire orbit, it is natural to change the coordinates from (x, y) to (h, s) according to the solution curve to (3.3), where we use s to denote time, $0 \le s \le \tau(h, \alpha)$. We have

$$\left(\frac{\partial x}{\partial t}\right)_{H=h} = x(1-y),\ \ \left(\frac{\partial y}{\partial t}\right)_{H=h} = \alpha y(x-1); \tag{3.37}$$

$$\left(\frac{\partial x}{\partial h}\right)_s\left(\alpha-\frac{\alpha}{x}\right)+\left(\frac{\partial y}{\partial h}\right)_s\left(1-\frac{1}{y}\right) = 1. \tag{3.38}$$

Therefore:

$$\det\left[\frac{D(x,y)}{D(h,s)}\right] = xy = G(x,y). \tag{3.39}$$

Then, the generalized relative entropy can be expressed as

$$\begin{aligned}
& \int_{\mathfrak{D}_h} u(x,y,t)\Psi\left(\frac{u(x,y,t)}{G^{-1}(x,y)\rho(H(x,y))}\right)\mathrm{d}x\mathrm{d}y \\
= & \int_{h_{min}}^{h}\mathrm{d}\eta\int_0^{\tau(\eta,\alpha)}\frac{u\big(x(s),y(s),t\big)}{G^{-1}\big(x(s),y(s)\big)}\Psi\left(\frac{u\big(x(s),y(s),t\big)}{G^{-1}\big(x(s),y(s)\big)\rho(\eta)}\right)\mathrm{d}s \\
= & \int_{h_{min}}^{h}\rho(\eta)\mathrm{d}\eta\int_0^{\tau(\eta,\alpha)}\widetilde{u}\big(x(s),y(s),t;\eta\big)\Psi\Big(\widetilde{u}\big(x(s),y(s),t;\eta\big)\Big)\mathrm{d}s \\
= & \int_{h_{min}}^{h}\Omega_B(\eta)\rho(\eta)\mathrm{d}\eta, \qquad (3.40)
\end{aligned}$$

in which

$$\widetilde{u}(x,y,t;h) = \frac{u\big(x,y,t\big)}{G^{-1}\big(x,y\big)\rho(h)}.$$

and

$$\Omega_B(h) = \oint_{\Gamma_{H=h}}\widetilde{u}\big(x,y,0;h\big)\Psi\Big(\widetilde{u}\big(x,y,0;h\big)\Big)\mathrm{d}\ell. \qquad (3.41)$$

$\Omega_B(h)$ is known as Boltzmann's entropy in classical statistical mechanics. We see that the $\mathcal{A}$ introduced in Sec. 3.1.2 is the simplest case of the generalized relative entropy in (3.36) with $\rho = \Psi = 1$, and $u(x,y,0) = G^{-1}(x,y)$. Then $\Omega_B(h) = \mathrm{d}\mathcal{A}_{\mathfrak{D}_h}/\mathrm{d}h$. Gibbs' canonical ensemble chooses $\rho(h) = e^{-h/\theta}$.

The dynamics (3.3) is not ergodic in the xy-plane; it does not have a unique invariant measure, as indicated by the arbitrary $\rho(H)$ in Eq. (3.34). However, the function $G(x,y)$, as indicated in Eqs. (3.10) and (3.39), is the unique invariant measure on each ergodic invariant submanifold $\Gamma_{H=h}$. It is non-uniform with respect to Lebesgue measure. On the ergodic invariant manifold $\Gamma_{H=h}$: $G(x,y)\mathrm{d}t \leftrightarrow \mathrm{d}\ell$. To see the difference between the Lebesgue-based average and invariant-measure based average, consider a simple time-varying exponentially growing population: $\frac{\mathrm{d}u(t)}{\mathrm{d}t} = r(t)u(t)$. The regular time average of the per capita growth rate is

$$\frac{1}{\tau}\int_0^{\tau}\frac{1}{u(t)}\frac{\mathrm{d}u}{\mathrm{d}t}\mathrm{d}t = \frac{1}{\tau}\ln\left(\frac{u(\tau)}{u(0)}\right) = \frac{1}{\tau}\int_0^{\tau}r(t)dt.$$

The Lebesgue-based average is an "average growth rate per average capita"

$$\frac{\displaystyle\int_0^{\tau}\frac{1}{u(t)}\frac{\mathrm{d}u}{\mathrm{d}t}\,u(t)\mathrm{d}t}{\displaystyle\int_0^{\tau}u(t)\mathrm{d}t} = \frac{\displaystyle\int_0^{\tau}r(t)u(t)\mathrm{d}t}{\displaystyle\int_0^{\tau}u(t)\mathrm{d}t}.$$

In cyclic population dynamics, this latter quantity corresponds to the $G\big(x(t),y(t)\big)$ weighted per capita growth rate or "kinetic energy"

$$\frac{\displaystyle\int_0^{\tau(h,\alpha)}\left(\frac{\mathrm{d}\ln(y)}{\mathrm{d}t}\right)G\big(x(t),y(t)\big)\mathrm{d}t}{\displaystyle\int_0^{\tau(h,\alpha)}G\big(x(t),y(t)\big)\mathrm{d}t}=\frac{\displaystyle\int_0^{\tau(h,\alpha)}-\left(\frac{\mathrm{d}\ln(x)}{\mathrm{d}t}\right)G\big(x(t),y(t)\big)\mathrm{d}t}{\displaystyle\int_0^{\tau(h,\alpha)}G\big(x(t),y(t)\big)\mathrm{d}t}$$

$$=\quad\frac{\displaystyle\int_0^{\tau(h,\alpha)}x\left(\frac{\partial H}{\partial x}\right)G\big(x(t),y(t)\big)\mathrm{d}t}{\displaystyle\int_0^{\tau(h,\alpha)}G\big(x(t),y(t)\big)\mathrm{d}t}=\frac{\displaystyle\int_0^{\tau(h,\alpha)}y\left(\frac{\partial H}{\partial y}\right)G\big(x(t),y(t)\big)\mathrm{d}t}{\displaystyle\int_0^{\tau(h,\alpha)}G\big(x(t),y(t)\big)\mathrm{d}t}. \tag{3.42}$$

3.1.4 Stochastic description of finite populations

In this section, we show that the conservative dynamics in (3.3) is an emergent caricature of a robust stochastic population dynamics. This material can be found in many texts, e.g., [86]. But for completeness, we shall give a brief summary.

Assume the populations of the prey and the predator, $M(t)$ and $N(t)$, reside in a spatial region of size Ω. The discrete stochastic population dynamics follows a two-dimensional, continuous time birth-death process with transition probability rate

$$\begin{aligned}&\mathrm{Pr}\Big\{M(t+\Delta t)=k,N(t+\Delta t)=\ell\,\Big|M(t)=m,N(t)=n\Big\}\\ =\;&\left(m\delta_{k,m+1}+\frac{1}{\Omega}mn\delta_{k,m-1}+\frac{\alpha}{\Omega}nm\delta_{\ell,n+1}+\alpha n\delta_{\ell,n-1}\right)\Delta t+o(\Delta t).\end{aligned} \tag{3.43}$$

The discrete stochastic dynamics has an invariant measure:

$$\mathrm{Pr}^{ss}\Big\{M=m,N=n\Big\}=\frac{1}{mn}. \tag{3.44}$$

Then

$$\begin{aligned}p_{m,n}(t+\Delta t)&=p_{m,n}(t)\left[1-\left(m+\frac{1}{\Omega}mn+\frac{\alpha}{\Omega}nm+\alpha n\right)\Delta t\right]\\&+p_{m-1,n}(t)\Big[(m-1)\Delta t\Big]+p_{m+1,n}(t)\left[\frac{1}{\Omega}(m+1)n\Delta t\right]\\&+p_{m,n-1}(t)\left[\frac{\alpha}{\Omega}(n-1)m\Delta t\right]+p_{m,n+1}(t)\Big[\alpha(n+1)\Delta t\Big].\end{aligned}$$

That is

$$\begin{aligned}\frac{p_{m,n}(t+\Delta t)-p_{m,n}(t)}{\Delta t} &= -m\Big[p_{m,n}(t)-p_{m-1,n}(t)\Big]-p_{m-1,n}(t)\\ &+ \frac{1}{\Omega}mn\Big[p_{m+1,n}(t)-p_{m,n}(t)\Big]+\frac{1}{\Omega}np_{m+1,n}(t)\\ &- \frac{\alpha}{\Omega}nm\Big[p_{m,n}(t)-p_{m,n-1}(t)\Big]-\frac{\alpha}{\Omega}mp_{m,n-1}(t)\\ &+ \alpha n\Big[p_{m,n+1}(t)-p_{m,n}(t)\Big]+\alpha p_{m,n+1}(t). \qquad (3.45)\end{aligned}$$

For a very large Ω, the population *densities* at time t can be approximated by continuous random variables as $X(t)=\Omega^{-1}M(t)$ and $Y(t)=\Omega^{-1}N(t)$. Then Eq. (3.45) becomes a partial differential equation by setting $x=m/\Omega$, $y=n/\Omega$, and $u(x,y,t)=p_{m,n}(t)/\Omega$:

$$\begin{aligned}\frac{\partial u}{\partial t} &= -x\frac{\partial u}{\partial x}+\frac{1}{2}\Omega^{-1}x\frac{\partial^2 u}{\partial x^2}-u+\Omega^{-1}\frac{\partial u}{\partial x}\\ &+ xy\frac{\partial u}{\partial x}+\frac{1}{2}\Omega^{-1}xy\frac{\partial^2 u}{\partial x^2}+yu+\Omega^{-1}y\frac{\partial u}{\partial x}\\ &- \alpha xy\frac{\partial u}{\partial y}+\frac{\alpha}{2}\Omega^{-1}xy\frac{\partial^2 u}{\partial y^2}-\alpha xu+\alpha\Omega^{-1}x\frac{\partial u}{\partial y}\\ &+ \alpha y\frac{\partial u}{\partial y}+\frac{\alpha}{2}\Omega^{-1}y\frac{\partial^2 u}{\partial y^2}+\alpha u+\alpha\Omega^{-1}\frac{\partial u}{\partial y}+o(\Omega^{-1}).\end{aligned}$$

Rearranging the terms and writing $\epsilon=\Omega^{-1}$, we can perform the Kramers-Moyal expansion to obtain:

$$\begin{aligned}\frac{\partial u(x,y,t)}{\partial t} &= \nabla\cdot\Big(\epsilon\mathbf{D}(x,y)\nabla u-\mathbf{F}(x,y)u\Big)+\frac{\epsilon}{2}\left((y+1)\frac{\partial u}{\partial x}+\alpha(x+1)\frac{\partial u}{\partial y}\right)+o(\epsilon)\\ &= \epsilon\sum_{\xi=x,y}\sum_{\zeta=x,y}\frac{\partial^2}{\partial\xi\partial\zeta}D_{\xi\zeta}(x,y)u(x,y,t)-\nabla\cdot\Big(\mathbf{F}(x,y)\,u\Big), \qquad (3.46)\end{aligned}$$

with drift $\mathbf{F}(x,y)=\big(f(x,y),g(x,y;\alpha)\big)^{\mathrm{T}}$ and symmetric diffusion matrix

$$\mathbf{D}(x,y)=\begin{pmatrix} D_{xx}(x,y) & D_{xy}(x,y)\\ D_{yx}(x,y) & D_{yy}(x,y)\end{pmatrix}=\frac{1}{2}\begin{pmatrix} x(1+y) & 0\\ 0 & \alpha y(x+1)\end{pmatrix}.$$

Eq. (3.46) should be interpreted as a Fokker-Plank equation for the probability density function $u(x,y,t)\mathrm{d}x\mathrm{d}y=\Pr\big\{x<X(t)\le x+\mathrm{d}x, y<Y(t)\le y+\mathrm{d}y\big\}$. It represents a

continuous stochastic process $\big(X(t),Y(t)\big)$ following Itô integral [3, 59, 86]:

$$\begin{aligned} \mathrm{d}X(t) &= X(1-Y)\mathrm{d}t + \epsilon^{\frac{1}{2}}\sqrt{X(1+Y)}\,\mathrm{d}W_1(t) \\ \mathrm{d}Y(t) &= \alpha Y(X-1)\mathrm{d}t + \epsilon^{\frac{1}{2}}\sqrt{\alpha Y(X+1)}\,\mathrm{d}W_2(t) \end{aligned} \tag{3.47}$$

It is important to recognize that in the limit of $\epsilon \to 0$, the dynamics described by Eq. (3.46) is reduced to that in Eq. (3.32), which is equivalent to Eq. (3.3) via the method of characteristics.

Potential-current decomposition

It can be verified that the stationary solution to Eq. (3.46) is actually $G^{-1}(x,y) = (xy)^{-1}$, which is consistent with the discrete case (cf. Eq. 3.44), and also a stationary solution to the Liouville equation Eq. (3.32).

As suggested in [177, **?**], the right-hand-side of Eq. (3.46) has a natural decomposition:

$$\begin{aligned} & \nabla\cdot\Big(\epsilon\mathbf{D}(x,y)\nabla u - \mathbf{F}(x,y)u\Big) + \frac{\epsilon}{2}\left((y+1)\frac{\partial u}{\partial x} + \alpha(x+1)\frac{\partial u}{\partial y}\right) \\ = {} & \nabla\cdot\Big[\epsilon\mathbf{D}(x,y)\nabla u - \Big(\mathbf{F}(x,y) - \epsilon\mathbf{D}(x,y)\nabla\ln G(x,y)\Big)u\Big] \\ = {} & \nabla\cdot\Big[\epsilon\mathbf{D}\Big(u\nabla\ln u + u\nabla\ln G\Big) - \mathbf{F}u\Big] \\ = {} & \epsilon\nabla\cdot\mathbf{D}u\nabla\Big(\ln\big(G\,u\big)\Big) - \nabla\cdot\big(\mathbf{F}\,u\big) \end{aligned} \tag{3.48}$$

in which the first term is a self-adjoint differential operator and the second is skew-symmetric [134]. The equation from the first line to the second uses the fact $\nabla\ln G = -\big(x^{-1},y^{-1}\big)$, thus $\mathbf{D}\nabla\ln G = -\frac{1}{2}\big((y+1),\alpha(x+1)\big)$. In terms of the stochastic differential equation in divergence form, this decomposition corresponds to:

$$\begin{pmatrix} \mathrm{d}X \\ \mathrm{d}Y \end{pmatrix} = -\epsilon\mathbf{D}\nabla\ln G + G\begin{pmatrix} -H_y \\ H_x \end{pmatrix} + \epsilon^{\frac{1}{2}}\sqrt{2\mathbf{D}}\begin{pmatrix} \mathrm{d}W_1(t) \\ \mathrm{d}W_2(t) \end{pmatrix}. \tag{3.49}$$

Under this non-Itô interpretation of the stochastic differential equation, the finite population with fluctuations (i.e., $\epsilon \neq 0$) is unstable when $x,y > 0$. The system behaves as an unstable focus as shown in Fig. 4.3. The eigenvalues at the fixed point $\big(1+\epsilon, 1-\epsilon\big)$ are $\pm i\sqrt{\alpha} + \frac{1}{2}\epsilon(\alpha+1)$, corresponding to the unstable nature of the stochastic system.

On the other hand, the potential-current decomposition reveals that the system (3.3) will be structurally stable in terms of the stochastic model: Any perturbation of the model system will yield corresponding conserved dynamics close to (3.3). The conservative ecology is a robust emergent phenomenon.

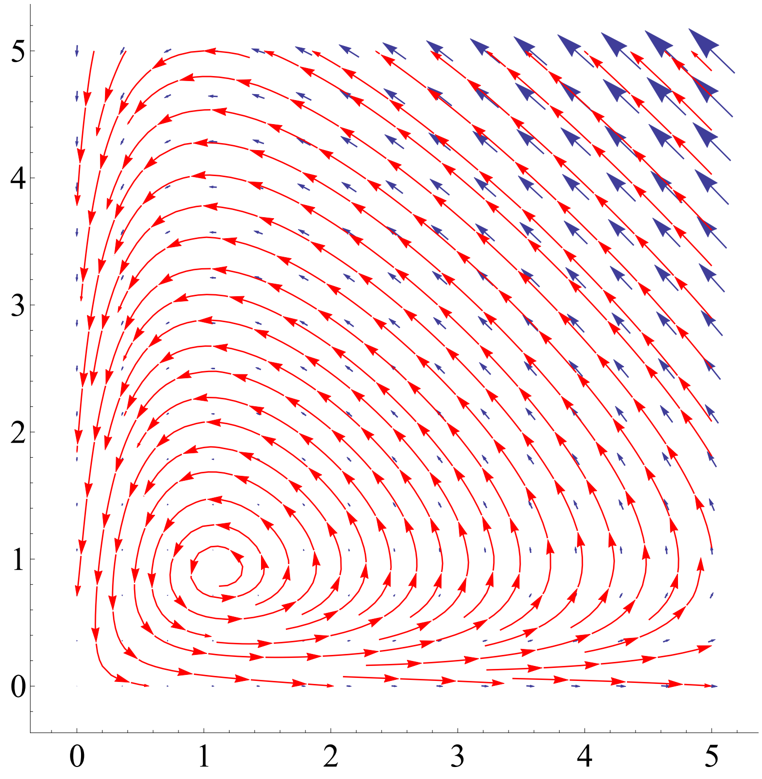


Figure 3.3: With fluctuation ($\epsilon = 0.1$), the deterministic part of system (3.49). Streamlines denote phase flow and arrows with different sizes denote the strength of the vector field.

Equations such as (3.47) and (3.49) are not mathematically well-defined until an precise meaning of integration

$$X(t) = \int_0^t b\big(W(t)\big)\mathrm{d}W(t) \tag{3.50}$$

is prescribed. This yields different stochastic processes $X(t)$ whose corresponding probability density function $f_{X(t)}(x,t)$ follow different linear partial differential equations. The fundamental solution to any partial differential equation (PDE), however, provides a Markov transition probability; there is no ambiguity at the PDE level. On the other hand, the only interpretation of (3.50) that provides a Markovian stochastic process that is non-anticipating is that of Itô's [51]. The differences in the interpretations of (3.50) become significant only in the modeling context, when one's intuition expects that $E\big[X(t)\big] = 0$ even for interpretations other then Itō's.

The slowly fluctuating $\boldsymbol{H_t = H\big(X(t), Y(t)\big)}$

With the $\big(X(t), Y(t)\big)$ defined in (3.46) and (3.47), let us now consider the stochastic functional

$$\begin{aligned} \mathrm{d}H\big(X(t), Y(t)\big) &= \alpha\left(1-\frac{1}{X}\right)\mathrm{d}X + \left(1-\frac{1}{Y}\right)\mathrm{d}Y + \frac{1}{2}\left(\frac{(\mathrm{d}X)^2}{X^2} + \frac{(\mathrm{d}Y)^2}{Y^2}\right) \\ &= \alpha\epsilon^{\frac{1}{2}}\left(\frac{(X-1)^2(1+Y)}{X} + \frac{(Y-1)^2(X+1)}{Y}\right)^{\frac{1}{2}}\mathrm{d}W(t) \\ &\quad + \frac{\epsilon}{2}\left(\frac{(1+Y)}{X} + \frac{\alpha(X+1)}{Y}\right)\mathrm{d}t \end{aligned} \tag{3.51}$$

Therefore, for very large populations, i.e., small ϵ, this suggests a separation of time scales between the cyclic motion on $\Gamma_{H=h}$ and slow, stochastic level crossing H_t. The method of averaging is applicable here [48, 190]:

$$\mathrm{d}H_t = \epsilon b(H_t)\mathrm{d}t + \epsilon^{\frac{1}{2}}A(H_t)\mathrm{d}W(t), \tag{3.52}$$

with

$$b(h) = \frac{1}{2}\left\langle\frac{(1+y)}{x} + \frac{\alpha(x+1)}{y}\right\rangle^{\Gamma_{H=h}}, \tag{3.53}$$

$$A(h) = \alpha\left\langle\left(\frac{(x-1)^2(1+y)}{x} + \frac{(y-1)^2(x+1)}{y}\right)^{\frac{1}{2}}\right\rangle^{\Gamma_{H=h}}, \tag{3.54}$$

where $\langle\psi(x,y)\rangle^{\Gamma_{H=h}} = \langle\psi(x,y)\rangle^{t}$ denotes the average of $\psi(x,y)$ on the level set $\Gamma_{H=h}$. Then, using the Itô integral, the distribution of H_t follows a Fokker-Planck equation:

$$\frac{\partial p(H,t)}{\partial t} = -\epsilon\frac{\partial}{\partial H}\big(b(H)p\big) + \frac{\epsilon}{2}\frac{\partial^2}{\partial H}\Big(A^2(H)p\Big). \tag{3.55}$$

And the stationary solution for Eq. (3.55) is:

$$p^{ss}(H) = \frac{1}{A^2(H)}\exp\left(2\int_{H_0}^{H}\frac{b(h)}{A^2(h)}\mathrm{d}h\right). \tag{3.56}$$

The steady state distributions of H under different α's are shown in Fig. 3.4. The steady state distribution $p^{ss}(H)$ does not depend on the volume size $\Omega = \epsilon^{-1}$.

When H is big enough, $p^{ss}(H)$ increases with H without bound, since $b(H)$ is a positive increasing function. Hence, $p^{ss}(H)$ is not normalizable on the entire $\mathbb{R}$, reflecting the unstable nature of the system. The fluctuation $A(H)$ approaches zero when H approaches $\alpha+1$. Consequently, the absorbing effect at $H = \alpha+1$ makes $p^{ss}(\alpha+1)$ another possible local maximum.

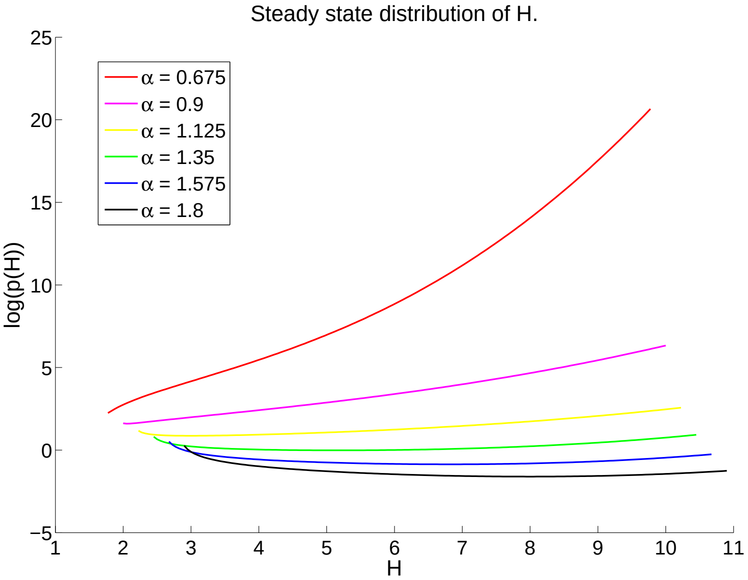


Figure 3.4: Under different values of α, the steady state distribution $p^{ss}(H)$ with respect to H in logarithmic scale. The slowly fluctuating "energy" H ranges from $\alpha+1$ to infinity. Its steady state distribution $p^{ss}(H)$ eventually increases without bound as H increases.

3.1.5 Discussion

It is usually an obligatory step in understanding an ODE $\dot{x} = f(x;\alpha,\beta)$ to analyze the dependence of a steady state x^* as an implicit function of the parameters (α,β) [109]. One of the important phenomena in this regard is the Thom-Zeeman catastrophe [109, 141]. From this broad perspective, the analysis developed by Helmholtz and Boltzmann in 1884 is an analysis of the geometry of a "non-constant but steady solution", as a function of its parameter(s) and initial conditions. In the context of LV equation (3.3), the geometry is characterized by the area encircled by a periodic solution, $\Gamma_{H=h}$, where h is specified by the the initial data: $\mathcal{A}\big(\mathfrak{D}_h\big) = \mathcal{A}\big(h,\alpha\big)$. The celebrated Helmholtz theorem [49, 25] then becomes our Eq. (3.25)

$$\mathrm{d}h = \frac{\mathrm{d}\mathcal{A} - \left(\dfrac{\partial\mathcal{A}}{\partial\alpha}\right)_h \mathrm{d}\alpha}{\left(\dfrac{\partial\mathcal{A}}{\partial h}\right)_\alpha} = \theta(h,\alpha)\,\mathrm{d}\ln\mathcal{A} - F_\alpha(h,\alpha)\mathrm{d}\alpha. \tag{3.57}$$

Since Eq. (3.3) has a conserved quantity H, Eq. (3.57) can, and should be, interpreted as an extended H conservation law, beyond the dynamics along a single trajectory, that includes both variations in α and in h. The partial derivatives in (3.57) can be shown as time averages of ecological activeness $\langle \ln x(\partial H/\partial \ln x)\rangle^t$ or $\langle \ln y(\partial H/\partial \ln y)\rangle^t$, and variation in the prey's population $\langle x - \ln x\rangle^t$. Those conjugate variables, along with parameter α, conserved quantity H, and encompassed area $\ln \mathcal{A}$ constitutes a set of "state variables" describing the state of an ecological system in its stationary, cyclic state. This is one of the essences of Boltzmann's statistical mechanics [49].

For the monocyclic Lotka-Volterra system, the dynamics are relatively simple. Hence, the state variables have monotonic relationships, the same as that observed in ideal gas models. When the system's dynamics become more complex (e.g. have more than one attractor, Hopf bifurcation), relations among the state variables will reflect that complexity (e.g. develop a cusp, exhibiting a phase transition in accordance [141]).

When the populations of predator and prey are finite, the stochastic predator-prey dynamics is unstable. This fact is reflected in the non-normalizable steady state distribution $G^{-1}(x, y)$ on $\mathbb{R}^{2+}$, and the destabilizing effect of the gradient dynamics in the potential-current decomposition. This is particular to the LV model we use; it is not a problem for the general theory if we study a more realistic model as in [185]. Despite the unstable dynamics, the stochastic model system is structurally stable: its dynamics persists under sufficiently small perturbations. This implies conservative dynamical systems like (3.3) are meaningful mathematical models, when interpreted correctly, for ecological realities.

Indeed, all ecological population dynamics can be represented by birth-death stochastic processes [86]. Except for systems with detailed balance, which rarely holds true, almost all such dynamics have underlying cyclic, stationary conservative dynamics. The present work shows that a hidden conservative ecological dynamics can be revealed through mathematical analyses. To recognize such a conservative ecology, however, several novel quantities need to be defined, developed, and becoming a part of ecological vocabulary. This is the intellectual legacy of Helmholtz's and Boltzmann's mechanical theory of heat [50].

3.2 Thermodynamics for General Stochastic Systems

We now turn our attention to the Gaussian fluctuation theory, one of the most successful branches of equilibrium statistical mechanics [90, 24], to develop a general thermodynamic theory. Since the work of Onsager and Machlup [117, 116], the Ornstein-Uhlenbeck process (OUP) has become the stochastic, mathematical description of dynamic, linear irreversible phenomena [180]. It has been extensively discussed in the literature in the past [32, 91, 92, 47]. Several recent papers studied particularly the OUP without detailed balance [131, 87]. In recent years, taking stochastic process rigorously developed by Kolmogorov as the mathematical representation, stochastic thermodynamics has emerged as the finite-time thermodynamic theory of mesoscopic systems, near and far from equilibrium [158, 173, 4]. The fundamental aspects of this new development are the mathematical notion of stochastic entropy production [142, 157, 52], novel thermodynamic relationships collectively known as nonequilibrium work equalities, and fluctuation theorems [74, 85, 33, 93, 64], and the mathematical concept of non-equilibrium steady-state [76, 189, 54].

Fundamental to all these advances is the notion of *time reversal*. Newtonian dynamic equation, in Hamiltonian form:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \frac{\partial H(x_i, y_i)}{\partial y_i}, \ \ \frac{\mathrm{d}y_i}{\mathrm{d}t} = -\frac{\partial H(x_i, y_i)}{\partial x_i}, \tag{3.58}$$

is a canonical example of dynamics with time-reversal symmetry [88]: Under transformation $(t, x_i, y_i) \longrightarrow (-t, x_i, -y_i)$, Eq. (3.58) is invariant. This invariance requires that $H(x_i, y_i) = H(x_i, -y_i)$: H is usually a function of y_i^2 and terms like $\vec{B} \cdot \vec{y}$, where $\vec{B}$ changes sign upon time reversal such as a magnetic field with a Lorentz force. Adopting this definition to linear stochastic processes, one has a novel definition for *time reversibility* that is distinctly different from that of Kolmogorov's, as we shall show below.

Consider the linear stochastic differential equation

$$\mathrm{d}\mathbf{X}(t) = -M(\alpha)\mathbf{X}(t)\mathrm{d}t + \epsilon\Gamma\mathrm{d}\mathbf{B}_t, \tag{3.59}$$

which is an OUP with parameters α and ϵ; M and Γ are two $n \times n$ constant matrices, $\mathbf{B}_t$ is standard Brownian motion. We further assume that all the eigenvalues of M are strictly positive and Γ is non-singular. According to the concept of detailed balance, Eq. (3.59) can be uniquely written as [174, 146, 9, **?**]

$$\mathrm{d}\mathbf{X}(t) = -D\Big\{\Xi^{-1} + \Big(D^{-1}M - \Xi^{-1}\Big)\Big\}\mathbf{X}\mathrm{d}t + \epsilon\Gamma\mathrm{d}\mathbf{B}_t, \tag{3.60a}$$

where D and $\Xi(\alpha)$ are positive definite matrices: $D = \frac{1}{2}\Gamma\Gamma^T$ and $M\Xi + \Xi M^T = 2D$. If one identifies the two terms inside $\{\cdots\}$ as dissipative (transient) and conservative (perpetuate) motions, respectively, then a time reversible process should be defined as a statistical equivalence between the probability density of a finite path $\{\mathbf{X}(t_0) = x_0, \mathbf{X}(t_1) = x_1, \cdots, \mathbf{X}(t_n) = x_n\}$ in which $t_0 < t_1 \cdots < t_n$:

$$f\big(x_0, x_1, \cdots, x_n\big),$$

and the probability density

$$f_{\mathbf{X}^\dagger(t_n)\mathbf{X}^\dagger(2t_n-t_{n-1})\cdots\mathbf{X}^\dagger(2t_n-t_0)}\big(x_n, x_{n-1}, \cdots, x_0\big)$$

in which the $\mathbf{X}^\dagger(t)$ follows the adjoint stochastic differential equation [**?**, 136]

$$\mathrm{d}\mathbf{X}^\dagger(t) = -D\Big\{\Xi^{-1} - \Big(D^{-1}M - \Xi^{-1}\Big)\Big\}\mathbf{X}^\dagger \mathrm{d}t + \epsilon\Gamma\mathrm{d}\mathbf{B}_t, \tag{3.60b}$$

with initial distribution for $\mathbf{X}^\dagger(t_n)$ identical to that of $\mathbf{X}(t_n)$.

Recognizing the underlying circulating, conservative dynamics in Eqs. (3.60a) and (3.60b) allows us to connect a Hamiltonian structure with linear stochastic processes, and consequently develop a Helmholtz theorem, which historically has served as the fundamental mathematical link between classical Newtonian mechanics and thermodynamics. For high dimensional stochastic processes, variables in the Helmholtz theorem provide the systems' underlying dynamics with a macroscopic picture. An ideal gas-like relation between a set of new, macroscopic variables emerges, confirming the simplicity of the OUP. A work-free energy equality in terms of the macroscopic thermodynamic variables, which are fluctuating with the underlying dynamics, captures the nature of the fluctuation in the underlying stochastic processes. We emphasize that even though the mathematical derivations are essentially the same, the physical meaning of the work relation is closer to the classical thermodynamics.

The subsections are structured as follows. In Sec. 3.2.1, we first provide the necessary preliminaries on the OUP. Sec. 3.2.1 introduces the conservative dynamics as a part of the stationary behavior of the OUP. Sec. 3.2.1 then discusses a long neglected issue of zero energy reference. Secs. 3.2.2 and 3.2.2 introduces the stationary free energy function and the dynamic free energy functional. Sec. 3.2.2 studies the novel object of equation of state. It is shown that the OUP has a simply, universal ideal thermodynamic behavior. In Sec. 3.2.3, we turn to the circulating dynamics and its relation to classical mechanics as well as stochastic dynamics. Sec. 3.2.3 focuses on the simplicity of the circulating dynamics as being totally integrable. Sec. 3.2.3 contains a proof that the stationary probability density of OUP, conditioned on an invariant torus of the underlying conservative dynamics, analogous to a microcanonical ensemble, is an invariant measure of the latter. If the dynamics on an invariant torus is ergodic, then the conditional probability is the only, natural invariant measure on the torus. Work equalities and fluctuation theorems are discussed in Sec. 3.2.4. Using a macroscopic presentation of the Jarzynski equality, its relation to Helmholtz theorem is revealed in Sec. 3.2.4. This section concludes with discussions in Sec. 3.2.5.

3.2.1 Preliminaries

Stationary Gaussian density and underlying conservative dynamics

The OUP in Eq. (3.60a) satisfies the important *fluctuation-dissipation relation*: $2D\Xi^{-1} = \epsilon^2\big(\Gamma\Gamma^T\big)\times$ covariance matrix of the stationary OUP. In fact, it has a stationary Gaussian distribution $Z^{-1}(\alpha)e^{-\varphi(\mathbf{x};\alpha)/\epsilon^2}$ in which $Z(\alpha)$ is a normalization factor and $\varphi(\mathbf{x};\alpha) = \mathbf{x}^T\Xi^{-1}(\alpha)\mathbf{x}$. In addition, there is an underlying circulating dynamics

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = -\Big(M(\alpha) - D\Xi^{-1}(\alpha)\Big)\mathbf{x}, \tag{3.61}$$

where the scalar $\varphi(\mathbf{x};\alpha)$ is conserved [87]:

$$\frac{\mathrm{d}}{\mathrm{d}t}\varphi\big(\mathbf{x}(t);\alpha\big) = -2\mathbf{x}^T\Xi^{-1}\Big(M - D\Xi^{-1}\Big)\mathbf{x} = -\mathbf{x}^T\Big(\Xi^{-1}M - M^T\Xi^{-1}\Big)\mathbf{x} = 0. \tag{3.62}$$

In fact, this conservative dynamics can be expressed as [**?**]:

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = -\frac{1}{2}\Big(M\Xi - D\Big)\nabla_{\mathbf{x}}\varphi(\mathbf{x};\alpha), \tag{3.63}$$

where

$$\varphi(\mathbf{x};\alpha) = \mathbf{x}^T\Xi^{-1}(\alpha)\mathbf{x}; \tag{3.64}$$

and $Q = \big(M(\alpha)\Xi - D\big)$ is skew-symmetric. This can be proved constructively as follows:

Proof 3 *For a linear system* $\dfrac{d\mathbf{x}}{dt} = U\mathbf{x}$*, since* $\phi(\mathbf{x}) = \dfrac{1}{2}\mathbf{x}^T\Xi^{-1}\mathbf{x}$ *is a conserved quantity,* $(\nabla\phi(\mathbf{x}))^TU\mathbf{x} = \mathbf{x}^T\Xi^{-1}U\mathbf{x} = 0$*.*

Since Ξ^{-1} *is positive definite, we have* $(\Xi^{-1}\mathbf{x})^TU\Xi(\Xi^{-1}\mathbf{x}) = 0$*, which means that* $Q = U\Xi$ *is skew symmetric and the dynamics can be written in terms of positive definite* Ξ^{-1} *and skew-symmetric* Q *as:* $\dfrac{d\mathbf{x}}{dt} = Q\Xi^{-1}\mathbf{x}$*.*

It is of paramount importance to recall that for a Markov process without detailed balance, its stationary dynamics is quantified by two mathematical objects: a stationary probability density and a stationary circulation [76, 177] characterized as a divergence-free, conservative vector field. In general, the latter accounts for the complexity arising from the system's dynamics [101]: how many integrals of motion does it have; whether the conservative dynamics is ergodic on an invariant set; etc. Many of the characteristics persist in the stationary stochastic process, and can be used to classify long time, complex behaviors in high dimensional systems. On the other hand, the dissipative (transient) dynamics plus noise drive the system towards the stationary distribution while characterizing "energy" fluctuations.

For the OUP in Eq. (3.60a), the conservative dynamics will be shown to be totally integrable. That is, symmetries would be implied through $\lfloor n/2 \rfloor$ first integrals of motions, which are the natural generalizations of the time-reversal symmetries. The remaining part, $\mathrm{d}\mathbf{X}(t) = -\frac{1}{2}D\nabla_{\mathbf{x}}\varphi(\mathbf{x};\alpha)\mathrm{d}t + \epsilon\Gamma\mathrm{d}\mathbf{B}_t$, has a stationary dynamics that is detailed balanced.

It is worth noting that any $\widetilde{\varphi}(\mathbf{x};\alpha) = \varphi(\mathbf{x};\alpha) + C(\alpha)$ is also a valid substitution for the $\varphi(\mathbf{x};\alpha)$ in Eq. (3.64). As far as the stochastic dynamical system Eq. (3.60a) is concerned, there is no unique $\widetilde{\varphi}(\mathbf{x};\alpha)$ as a function of both dynamic variable $\mathbf{x}$ and parameter α.

Zero energy reference: A hidden assumption in classical physics

The central object that connects classical Newtonian mechanics with equilibrium thermodynamics is the entropy function $S(E,V,N)$, with V and N being the volume and the number of particles of a classical mechanical system in a container, and E its total mechanical energy which is conserved according to Newton's Second Law of motion. In Hamilton's formulation Eq. (3.58), E is simply the initial value of the Hamiltonian function $H\big(\{x_i\},\{y_i\}\big)$ in which x_i and y_i are the position and momentum of ith particle, respectively, $1 \le i \le N$.

We recognize that in the classical theory of mechanical motions, replacing H with $\widetilde{H}\big(\{x_i\},\{y_i\}\big) = H\big(\{x_i\},\{y_i\}\big) + C$, where C is a constant, has absolutely no consequence to the mathematical theory. Therefore, with parameters contained in the Hamiltonian function, such as V and N, $H\big(\{x_i\},\{y_i\};V,N\big)$ and $H\big(\{x_i\},\{y_i\};V,N\big)+C(V,N)$ are equivalent. In other words, classical mechanics only uniquely determines a Hamiltonian function up to an arbitrary function of all the non-dynamic parameters.

However, an additive function $C(V,N)$ would cause non-uniqueness in the thermodynamic forces in the relation:

$$\mathrm{d}S(E,V,N) = \left(\frac{\partial S}{\partial E}\right)_{V,N}\Big[\mathrm{d}E + p\mathrm{d}V - \mu\mathrm{d}N\Big], \tag{3.65}$$

in which

$$p = \frac{\left(\dfrac{\partial S}{\partial V}\right)_{E,N}}{\left(\dfrac{\partial S}{\partial E}\right)_{V,N}} = -\left(\frac{\partial E}{\partial V}\right)_{S,N}, \quad \mu = \left(\frac{\partial E}{\partial N}\right)_{S,V}. \tag{3.66}$$

Corresponding to $\widetilde{H} = H + C(V, N)$ one has, for E as the initial values of $H\big(\{x_i\}, \{y_i\}; V, N\big)$, and $\widetilde{E}$ as the initial values of $\widetilde{H}\big(\{x_i\}, \{y_i\}; V, N\big)$:

$$\widetilde{p} = -\left(\frac{\partial \widetilde{E}}{\partial V}\right)_{S,N} = p - \left(\frac{\partial C}{\partial V}\right)_N, \;\; \widetilde{\mu} = \mu + \left(\frac{\partial C}{\partial N}\right)_V. \tag{3.67}$$

Since pressure p has a mechanical interpretation, one can, by physical principle, uniquely determine the form of p as a function of V. The situation for μ is much less clear: Since there is not an independent mechanical interpretation of the chemical potential other than the thermodynamic one given in Eq. (3.65), the non-uniqueness is inherent in the mathematical, as well as the physico-chemical theory. The problem has the same origin as *Gibbs' paradox* [114, 53].

In classical chemical thermodynamics, the Hamiltonian function as a function of varying number of particles N, $H(x_1, \cdots, x_N, y_1, \cdots, y_N) = \frac{1}{2}\sum_i^N m_i y_i^2 + V(x_1, \cdots, x_N)$, is uniquely determined via a Kirkwood charging process [125]:

$$V\big(x_1, \cdots, x_N, x_{N+1}\big) = V\big(x_1, \cdots, x_N\big) + \sum_{j=1}^{N} U_{j,N+1}\big(x_j, x_{N+1}\big), \tag{3.68}$$

where

$$\lim_{|x-y|\to\infty} U(x, y) = 0. \tag{3.69}$$

With this convention, the Hamiltonian for a molecular system is uniquely determined in chemical thermodynamics, which yields a consistent chemical potential μ. How to generalize this chemical approach to Hamiltonian dynamics Eq. (3.58) with no clear separation between kinetic and potential parts, however, is unclear.

The problem of uniqueness of Hamiltonian function $\widetilde{H}$ is intimately related to the uniqueness of $\widetilde{\varphi}(\mathbf{x}; \alpha)$ in Sec. 3.2.1. As we shall show in the rest of this section, the zero energy reference has deep implications to the theory of stochastic thermodynamics. The resolution to the problem will be discussed in Sec. 3.2.3.

3.2.2 Free energy functions and functional

As the notion of entropy, the definition of free energy is widely varied in the literature.[1] The most general features of free energy, perhaps, are: it is the difference between "internal

[1] It has become increasingly clear that the Boltzmann's entropy for a Hamiltonian dynamics is not unique: There are different geometric characterizations of the level sets of the Hamiltonian that can be acceptable choices. Neither is Shannon's entropy in stochastic dynamics unique: other convex functions such as Tsallis' entropy can also be found in the literature.

energy" and entropy; it is the entropy under a "natural invariant measure". In this section, we shall present two different types of free energies associated with the OU dynamics in Eq. (3.60a):

(i) Thermodynamic free energy of a stationary dynamics, as a function of mean internal energy E and parameter α: $A(E, \alpha)$. We identify a "thermodynamic state" as a state of sustained motion, either for a deterministic conservative dynamics Eq. (3.61), or for a stochastic stationary process defined by Eq. (3.60a).

(ii) Dynamic free energy functional, $\Psi[f(\mathbf{x}, t)]$, for an instantaneous probability distribution $f(\mathbf{x}, t)$.

Thermodynamic free energy functions $A(E, \alpha)$

With a particular given $\varphi(\mathbf{x}; \alpha)$, we now introduce two different free energy functions. The first one is defined following the microcanonical ensemble approach; definition of the second one follows Gibb's canonical ensemble approach. While the second one is frequently being used in the work-free energy relation (discussed in Sec. 5), the two definitions agree perfectly in the large dimension limit.

The first thermodynamic free energy function, $A_1(E, \alpha)$, associated with the conservative deterministic motion of Eq. (3.61) on the surface of $\varphi(\mathbf{x}; \alpha) = E$, is obtained following the microcanonical ensemble approach through Boltzmann's entropy function. Letting $\sigma_B(E, \alpha)$ correspond to the entropy S and $\Theta^{-1}(E, \alpha)$ correspond to $\left(\frac{\partial S}{\partial E}\right)$ in Eq. (3.65), we can define:

$$\begin{aligned}
\sigma_B(E, \alpha) &= \ln\left(\int_{\varphi(\mathbf{x};\alpha)\le E} \mathbf{dx}\right) \\
&= \frac{n}{2}\ln E + \frac{1}{2}\ln\det\Xi(\alpha) + \ln V_n, && \text{(3.70a)} \\
\Theta^{-1}(E, \alpha) &= \left(\frac{\partial \sigma_B}{\partial E}\right)_{\alpha} = \frac{n}{2E}, && \text{(3.70b)} \\
A_1(E, \alpha) &= E - \Theta\sigma_B \\
&= \frac{2E}{n}\left\{-\frac{n}{2}\ln E - \frac{1}{2}\ln\det\Xi(\alpha) - \frac{n}{2}\ln(\pi) + \ln\Gamma\left(\frac{n}{2}+1\right) + \frac{n}{2}\right\}, && \text{(3.70c)}
\end{aligned}$$

where $V_n = \pi^{n/2}\left(\Gamma\left(\frac{n}{2}+1\right)\right)^{-1}$ is the volume of an n-dimensional Euclidean ball with radius 1. $\Gamma(\cdot)$ is gamma function. n is the dimension of the OUP in Eq. (3.60a).

The second one, $A_2(E, \alpha)$, follows Gibbs' canonical ensemble approach via the "partition function" $Z(\alpha)$:

$$\begin{aligned} Z(\alpha) &= \int_{\mathbb{R}^n} e^{-\varphi(\mathbf{x};\alpha)/\epsilon^2}\mathbf{dx} \;=\; \Big(\big(\pi\epsilon^2\big)^n \det\Xi(\alpha)\Big)^{\frac{1}{2}}, && (3.71\text{a}) \\ A_2(\overline{E},\alpha) &= -\epsilon^2 \ln Z(\alpha) \\ &= \frac{2\overline{E}}{n}\left\{-\frac{n}{2}\ln\overline{E} - \frac{1}{2}\ln\det\Xi(\alpha) - \frac{n}{2}\ln(\pi) + \frac{n}{2}\ln\Big(\frac{n}{2}\Big)\right\}, && (3.71\text{b}) \end{aligned}$$

in which mean internal energy

$$\overline{E} = \frac{1}{Z(\alpha)}\int_{\mathbb{R}^n}\varphi(\mathbf{x};\alpha)e^{-\varphi(\mathbf{x};\alpha)/\epsilon^2}\mathbf{dx} = \frac{1}{2}n\epsilon^2. \qquad (3.71\text{c})$$

The two free energy functions A_1 in Eq. (3.70c) and A_2 in Eq. (3.71b) are different only by a function of n inside the $\{\cdots\}$. For large n, $\ln\Gamma(\frac{n}{2}+1) \approx \frac{n}{2}\ln\left(\frac{n}{2}\right) - \frac{n}{2}$. Therefore, A_1 and A_2 agree perfectly in the limit of $n \to \infty$.

Dynamic free energy functional $\Psi[f_\alpha(\mathbf{x},t)]$

The thermodynamic free energy $A_2(E,\alpha)$ in Sec. 3.2.2 sets a universal energy reference point for the entire family of stochastic dynamics in Eq. (3.60a) with different α. For a given α, the time-dependent probability density function $f_\alpha(\mathbf{x},t)$ follows the partial differential equation

$$\frac{\partial f_\alpha(\mathbf{x},t)}{\partial t} = \nabla_{\mathbf{x}}\cdot\Big(\epsilon^2 D\nabla_{\mathbf{x}} f_\alpha(\mathbf{x},t) + M(\alpha)\mathbf{x}f_\alpha(\mathbf{x},t)\Big). \qquad (3.72)$$

The $f_\alpha(\mathbf{x},t)$ represents an instantaneous "state" of the probabilistic system, which has a free energy functional

$$\begin{aligned} \Psi\big[f_\alpha(\mathbf{x},t)\big] &= \int_{\mathbb{R}^n}\varphi(\mathbf{x};\alpha)f_\alpha(\mathbf{x},t)\,\mathbf{dx} - \left(-\epsilon^2\int_{\mathbb{R}^n} f_\alpha(\mathbf{x},t)\ln f_\alpha(\mathbf{x},t)\,\mathbf{dx}\right) \\ &= \epsilon^2\int_{\mathbb{R}^n} f_\alpha(\mathbf{x},t)\ln\left(\frac{f_\alpha(\mathbf{x},t)}{Z^{-1}(\alpha)e^{-\varphi(\mathbf{x};\alpha)/\epsilon^2}}\right)\mathbf{dx} + A_2(\alpha). && (3.73) \end{aligned}$$

This is a dynamic generalization of the free energy functions in Sec. 3.2.2. It has two important properties. First,

$$\lim_{t\to\infty}\Psi\big[f_\alpha(\mathbf{x},t)\big] = A_2(\alpha). \qquad (3.74)$$

Second [**?**, 136],

$$\frac{\mathrm{d}}{\mathrm{d}t}\Psi\big[f_\alpha(\mathbf{x},t)\big] \le 0, \qquad (3.75)$$

in which the equality holds if and only if $f_\alpha(\mathbf{x},t)$ reaches its stationary distribution $Z^{-1}(\alpha)e^{-\varphi(\mathbf{x};\alpha)/\epsilon^2}$. The negated rate of change in the dynamic free energy functional, $-\mathrm{d}\Psi/\mathrm{d}t$, is widely recognized as non-adiabatic entropy production rate.

The entropy production rate also has a finite time, stochastic counterpart in terms of the logarithm of the likelihood ratio:

$$-\frac{\mathrm{d}}{\mathrm{d}t}\Psi\big[f_\alpha(\mathbf{x},t)\big] = \lim_{s\to t} E^{\mathbb{P}}\left[\frac{1}{|s-t|}\ln\left(\frac{f\big(\mathbf{X}(\tau)|t\le\tau\le s\big)}{f_{\mathbf{X}^\dagger}\big(\hat{\mathbf{X}}(\tau)|t\le\tau\le s\big)}\right)\right], \tag{3.76}$$

where $\hat{\mathbf{X}}(\tau)=\mathbf{X}(s-\tau+t)$, and the expectation $E^{\mathbb{P}}\big[\cdots\big]$ is carried out over the diffusion process defined by Eq. (3.60a) and the corresponding Eq. (3.72):

$$f\big(\mathbf{X}(\tau)|t\le\tau\le s\big) \propto \exp\left[-2\int_t^s \Big(\dot{\mathbf{X}}(\tau)+M\mathbf{X}(\tau)\Big)^T D\Big(\dot{\mathbf{X}}(\tau)+M\mathbf{X}(\tau)\Big)\mathrm{d}\tau\right]. \tag{3.77}$$

Universal equation of state of OU process

With the introduction of the internal energy E and the parameter α, the thermodynamic relation Eq. (3.65) - known as the Helmholtz theorem - for the OUP model can be expressed by σ_B, α and their conjugate variables. We notice that α enters Eq. (3.70) only through $\det\Xi(\alpha)$. If one measures α through $\widetilde{\alpha}=\det\Xi(\alpha)$, then the Helmholtz theorem writes:

$$\begin{aligned}\mathrm{d}E &= \Theta(E,\widetilde{\alpha})\mathrm{d}\sigma_B - F_{\widetilde{\alpha}}(E,\widetilde{\alpha})\mathrm{d}\widetilde{\alpha}\\ &= \left(\frac{\partial\sigma_B}{\partial E}\right)^{-1}\mathrm{d}\sigma - \left(\frac{\partial\sigma_B}{\partial\widetilde{\alpha}}\right)\left(\frac{\partial\sigma_B}{\partial E}\right)^{-1}\mathrm{d}\widetilde{\alpha}.\end{aligned} \tag{3.78}$$

The two conjugate variables, Θ and $F_{\widetilde{\alpha}}$, correspond to the macroscopic quantities in classical thermodynamics as temperature and force.[2]

Following either Boltzmann's microcanonical or Gibbs' canonical approach, Sec. 3.2.2 revealed that $E=\frac{1}{2}n\Theta$ in which $\theta=\frac{1}{2n}\Theta$ could be interpreted as an "absolute temperature". Since the absolute temperature θ is a fluctuating quantity with respect to E and $\widetilde{\alpha}$, it may, in general, not bear a simple relationship with the noise strength ϵ^2. But here in OUP, by comparing the microcanonical approach with the canonical one, we note that the mean absolute temperature $\bar{\theta}=\frac{\epsilon^2}{2n}$.

[2] The force here should be understood as Onsager's thermodynamic force: corresponding to a spatial displacement is a mechanical force; to a change in number of particles is Gibbs' chemical potential; to a variation in a parameter through a Maxwell demon then is an informatic force [168, 106].

The thermodynamic conjugate variable of $\widetilde{\alpha}$, the $\widetilde{\alpha}$-force:

$$F_{\widetilde{\alpha}} = \Theta \left(\frac{\partial \sigma_B(E, \widetilde{\alpha})}{\partial \widetilde{\alpha}} \right)_E = \frac{n\theta}{\widetilde{\alpha}}. \tag{3.79}$$

A mathematical relation between $\widetilde{\alpha}$, $F_{\widetilde{\alpha}}$, and θ is called an *equation of state* in classical thermodynamics.

The "internal energy" E being a sole function of temperature θ, and the product of thermodynamic conjugate variables, $\widetilde{\alpha}F_{\widetilde{\alpha}}$, equaling to $n\theta$, are hallmarks of thermodynamic behavior of ideal gas and ideal solution. We thus conclude that the OUP has a universal ideal thermodynamic behavior.

3.2.3 Circulating conservative flow and its invariant measures

After discussing the energy function and stationary probability, we now focus on the dynamic complexity of the system and study the circulating, conservative dynamics. The universal ideal thermodynamic behavior reveals one aspect of the simplicity in OUP; another is reflected in the divergence-free motions. For the linear conservative dynamics, Eq. (3.61), its structure is known to be simple: the vector field is integrable.

The conservative dynamics in Eq. (3.61) can be proved to be purely cyclic (e.g., periodic, or quasi-periodic on an invariant torus). Because matrix $Q = M\Xi - D$ is skew-symmetric and hence $(M - D\Xi^{-1}) = Q\Xi^{-1}$ has only pairs of imaginary eigenvalues $\{\lambda_\ell | 1 \le \ell \le n\}$.

Proof 4 *Since U is positive definite, we can write its eigendecomposition as: $U = V_U \Lambda_U V_U^*$, where Λ_U is a diagonal matrix with only positive terms on diagonal; and since Q is skew-symmetric, it is also unitarily diagonalizable: $Q = V_Q \Lambda_Q V_Q^*$.*

Take $\hat{\mathbf{x}} = \sqrt{\Lambda_U}\ V_U^*\ \mathbf{x}$, and write $\tilde{V} = V_U^*\ V_Q$, then:

$$\frac{d\hat{\mathbf{x}}}{dt} = \sqrt{\Lambda_U}\ V_U^*\ \frac{d\mathbf{x}}{dt} = \sqrt{\Lambda_U}\ V_U^*\ QU\mathbf{x} = \left(\sqrt{\Lambda_U}\ \tilde{V}\right)\ \Lambda_Q \left(\tilde{V}^* \sqrt{\Lambda_U}\right) \hat{\mathbf{x}} = A\ \hat{\mathbf{x}}. \tag{3.80}$$

Since Λ_U is a diagonal matrix with only positive terms on diagonal, $\sqrt{\Lambda_U}$ is a real diagonal matrix: $\left(\sqrt{\Lambda_U}\right)^* = \sqrt{\Lambda_U}$, $\left(\tilde{V}^* \sqrt{\Lambda_U}\right) = \left(\sqrt{\Lambda_U}\ \tilde{V}\right)^*$.

Hence, QU is similar to the matrix $A = \left(\sqrt{\Lambda_U}\ \tilde{V}\right)\ \Lambda_Q \left(\sqrt{\Lambda_U}\ \tilde{V}\right)^*$. Note that:

$$A^* = \left(\sqrt{\Lambda_U}\ \tilde{V}\right)\ \Lambda_Q^* \left(\sqrt{\Lambda_U}\ \tilde{V}\right)^* = -\left(\sqrt{\Lambda_U}\ \tilde{V}\right)\ \Lambda_Q \left(\sqrt{\Lambda_U}\ \tilde{V}\right)^* = -A. \tag{3.81}$$

Therefore, A is skew-Hermitian, having purely imaginary eigenvalues. Matrix QU is similar to A. Hence QU have only imaginary eigenvalues. Therefore, the motion described by Eq. 1.2 is purely cyclic on an invariant torus.

We can also find real Jordan form of $(M - D\Xi^{-1})$: PJP^{-1}, where J is block diagonal, with 2×2 skew-symmetric blocks:

$$\mathrm{Im}\left[\lambda_{(2i-1)}\right]\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

being the ith block on the diagonal. Natural coordinates for the conservative flow Eq. (3.61) is therefore: $\mathbf{y} = P^{-1}\mathbf{x}$.

The conservative flow and general time reversal symmetries

Poisson bracket $\{\cdot,\cdot\}$ can be defined for the linear conservative system as: $\{\varphi(\mathbf{x}), \psi(\mathbf{x})\} = \nabla\varphi(\mathbf{x})^T(M - D\Xi^{-1})\nabla\psi(\mathbf{x})$. Then the conservative flow expressed in terms of its Hamiltonian function $\varphi(\mathbf{x})$ is:

$$\dot{x_i} = \big\{x_i, \frac{1}{2}\varphi(\mathbf{x})\big\}. \tag{3.82}$$

First integrals I_i of the conservative flow are:

$$I_i = y_{2i-1}^2 + y_{2i}^2 = \mathbf{x}^T P^{-T} I_{(2i-1)\sim(2i)} P^{-1}\mathbf{x},\ \ 1 \le i \le \left\lfloor \frac{n}{2} \right\rfloor. \tag{3.83}$$

Here, $I_{(2i-1)\sim(2i)}$ denotes the diagonal matrix with 1 on $(2i-1)$-th to $(2i)$-th diagonal entries, and zero everywhere else.

The conservative flow is totally integrable, and can be written in canonical action-angle variables. Angular coordinates θ_i accompanying I_i can be found as:

$$\theta_i = \mathrm{Im}\big[\lambda_{2i-1}\big]^{-1} \cdot \arctan\left(\frac{y_{2i-1}}{y_{2i}}\right),\ \ 1 \le i \le \left\lfloor \frac{n}{2} \right\rfloor. \tag{3.84}$$

Hence, in the canonical action-angle variables, $\varphi = \sum\limits_{i=1}^{\lfloor n/2 \rfloor} I_i$,

$$\begin{cases} \theta_i' = \dfrac{\partial\varphi}{\partial I_i} = 1 \\ I_i' = -\dfrac{\partial\varphi}{\partial\theta_i} = 0. \end{cases} \tag{3.85}$$

There are $\left\lfloor \frac{n}{2} \right\rfloor$ first integrals, but for the given Poisson bracket, one combination of them is unique, which is the Hamiltonian φ that connects to the stationary distribution and generates the conservative flow.

In the action-angle variables, it is observable that the system bears the following symmetries: $(t, \theta_i, I_i) \longrightarrow \big(-t, -\theta_i, (-1)^{k_i} I_i\big)$, where $\{k_i\}$ is a sequence of 0 and 1. Taking $\{k_i\}$ as a sequence of zeros, we recover the time-reversal invariance in Eq. (3.58). Hence, for general $\{k_i\}$, those symmetries are the natural generalizations of the time-reversal symmetry, as displayed in classical Hamiltonian systems.

Conditional probability measure as invariant measure of the conservative flow

The OUP yields an equilibrium probability density function for $\mathbf{X}$:

$$f_{\mathbf{X}}^{eq}(\mathbf{x};\alpha) = Z^{-1}(\alpha) e^{-\varphi(\mathbf{x};\alpha)/\epsilon^2}.$$

In this section, we calculate the *conditional probability density* for $\mathbf{X}$ restricted on an equal energy surface $\mathfrak{D}_{\varphi=E} = \big\{\mathbf{x}|\mathbf{x} \in \mathbb{R}^n, \varphi(\mathbf{x};\alpha) = E\big\}$ and prove it to be an invariant measure of the conservative dynamics, Eq. (3.61), restricted on $\mathfrak{D}_{\varphi=E}$. Therefore, in the absence of fluctuation and dissipation, our definition of "equilibrium free energy" (Eq. (3.70c)) in stochastic thermodynamics retreats to the Boltzmann's microcanonical ensemble approach in classical mechanics.

One can obtain a conditional probability density for $\mathbf{X}$ restricted on an equal energy surface $\mathfrak{D}_{\varphi=E} = \big\{\mathbf{x}|\mathbf{x} \in \mathbb{R}^n, \varphi(\mathbf{x};\alpha) = E\big\}$ as:

$$Z(\alpha) = \int_{\mathbb{R}^n} e^{-\varphi(\mathbf{x};\alpha)/\epsilon^2} \mathrm{d}\mathbf{x} = \int_{\varphi_{min}(\alpha)}^{\infty} \exp\left(-\frac{E}{\epsilon^2} + S(E,\alpha)\right) \mathrm{d}E, \tag{3.86}$$

in which [72]

$$S(E,\alpha) = \ln\left(\frac{\partial}{\partial E}\int_{\varphi(\mathbf{x};\alpha)\le E} \mathrm{d}\mathbf{x}\right)_{\alpha} = \ln\left(\oint_{\varphi(\mathbf{x};\alpha)=E} \frac{\mathrm{d}\Sigma^{n-1}}{\|\nabla_{\mathbf{x}}\varphi(\mathbf{x};\alpha)\|}\right). \tag{3.87}$$

The conditional probability density at $\mathbf{x} \in \mathfrak{D}_{\varphi=E}$ is:

$$\frac{e^{-S(E,\alpha)}}{\|\nabla_{\mathbf{x}}\varphi(\mathbf{x};\alpha)\|} = \frac{1}{\frac{n}{2} V_n E^{\frac{n}{2}-1}\big(\det \Xi\big)^{-\frac{1}{2}} \|\Xi^{-1}\mathbf{x}\|}. \tag{3.88}$$

Note this conditional probability is one of the invariant measures of the conservative dynamics Eq. (3.61) restricted in $\mathfrak{D}_{\varphi=E}$.

To prove this fact, define the dynamics of the conservative part as: $\mathbf{S}_t$, mapping a measurable set $A \to \mathbf{S}_t(A)$. Then measure of a set $A \subseteq \mathfrak{D}_{\varphi=E}$ under

$$\mathrm{d}\mu = e^{-S(E,\alpha)} \|\nabla_{\mathbf{x}}\varphi(\mathbf{x};\alpha)\|^{-1} \mathrm{d}\Sigma^{n-1}$$

is:

$$\int_A \frac{e^{-S(E,\alpha)}}{\|\nabla_{\mathbf{x}}\varphi(\mathbf{x};\alpha)\|}\mathrm{d}\Sigma^{n-1} = e^{-S(E,\alpha)}\int_A \delta(E-\varphi(\mathbf{x}))\mathrm{d}\mathbf{x}$$
$$= e^{-S(E,\alpha)}\int_{\mathbf{S}_t^{-1}(A)} \delta(E-\varphi(\mathbf{x}))\mathrm{d}\mathbf{x}$$
$$= \int_{S_t^{-1}(A)} \frac{e^{-S(E,\alpha)}}{\|\nabla_{\mathbf{x}}\varphi(\mathbf{x};\alpha)\|}\mathrm{d}\Sigma^{n-1}, \qquad (3.89)$$

since $\mathbf{S}_t^{-1}(A) \subseteq \mathfrak{D}_{\varphi=E}$ and $\mathbf{S}_t$ is volume preserving. In general, if the dynamics is ergodic on the entire $\mathfrak{D}_{\varphi=E}$, then its invariant measure μ is the *physical measure*: μ-average equals time average along a trajectory; if there are other first integrals for the conservative dynamics, then μ can be projected further to lower dimensional invariant sets.

Resolutions to the energy reference problem

Up to now, there are clearly several possibilities to uniquely determine the free additive function $C(V,N)$ in the Hamiltonian $H\big(\{x_i\},\{y_i\};V,N\big)$ discussed in Introduction.

(1) $C(V,N)$ is chosen such that the global minimum of $H=0$ for each and every V and N. This is widely used, implicitly, in application practices, as in our Eq. (3.64).

(2) $C(V,N)$ is chosen according to the "equilibrium free energy":

$$-\epsilon^2 \ln \int e^{-E/\epsilon^2 + S(E,V,N)}\mathrm{d}E = 0. \qquad (3.90)$$

Note that this is precisely the "energy function" in Hatano and Sasa [64].

(3) Extra information concerning the fluctuations in V, such as in an isobaric ensemble, and fluctuations in N in grand ensemble, provides an empirically determined basis for the free energy scale.

In terms of the theory of probability, choice (2) uniquely determines the energy reference point according to a conditional probability, and in choice (3) it is uniquely determined according to a marginal distribution. How to normalize a probability, which has always been considered non-consequential in statistical physics, seems to be a fundamental problem in the physics of complex systems.

3.2.4 Work equalities and fluctuation theorems

The previous discussions suggest that while a great deal of complexity of a detailed, mesoscopic stationary dynamics is captured by the circulating conservative dynamics, OUP also has a macroscopic state of motion that is defined by the internal energy E, or equivalently the level sets of $\varphi(\mathbf{x})$,$\mathfrak{D}_{\varphi=E} \subset \mathbb{R}^n$. Thus, from the macroscopic point of view, a stochastic system could be studied through the one-dimensional (1-D) time sequence of fluctuating internal energy E, as a function of t, or the change in E due to changes in the parameter α.

The celebrated Jarzynski equality connects the mesoscopic fluctuating force with the change in free energy. We present this result through a projection from n-D phase space to 1-D function $E(t)$ that facilitates experimental verification of the work-free energy relation. This approach reveals a close connection between the Jarzynski equality and the Helmholtz theorem. We start with stating the Jarzynski and Crooks' equalities, with the mathematical proofs collected here. We then demonstrate the novel formulation of the Jarzynski equality in the projected space.

We have shown that $A_2(\alpha)$ is uniquely determined only up to a particular $\varphi(\mathbf{x};\alpha)$. As shown below, the existence of an $A_2(\alpha)$ has a paramount importance in the theories of work equalities, in which the notion of a common energy for a family of stochastic dynamical systems with different α has to be given *a priori* [159].

The Jarzynski equality

The macroscopic α-force in the Helmholtz theorem, as a function of E and α, is defined through Boltzmann's entropy σ_B. The Jarzynski equality, on the other hand, concerns with a mesoscopic α-force,

$$F_\alpha(\mathbf{x};\alpha) = -\left(\frac{\partial\varphi(\mathbf{x};\alpha)}{\partial\alpha}\right)_{\mathbf{x}}, \tag{3.91}$$

and the statistical behavior of its corresponding *stochastic work*

$$W[\mathbf{X}(\tau),\alpha(\tau)] = \int_0^t F_\alpha\big(\mathbf{X}(\tau);\alpha(\tau)\big)\left(\frac{\mathrm{d}\alpha(\tau)}{\mathrm{d}\tau}\right)\mathrm{d}\tau. \tag{3.92}$$

The Jarzynski equality dictates that if the initial distribution of $\mathbf{X}(\tau)$ follows the equilibrium distribution, then [74]

$$\left\langle e^{-\frac{1}{\epsilon^2}W[\mathbf{X}(\tau),\alpha(\tau)]}\right\rangle_{\big[\mathbf{X}(\tau),\alpha(\tau)\big]} = e^{-\frac{1}{\epsilon^2}\Delta A_2(\alpha)}, \tag{3.93}$$

where the average of a functional over the ensemble of paths is defined as:

$$\Big\langle G[\mathbf{X}(\tau),\alpha(\tau)]\Big\rangle_{\big[\mathbf{X}(\tau),\alpha(\tau)\big]} = \int G[\mathbf{X}(\tau),\alpha(\tau)]\mathcal{P}[\mathbf{X}(\tau),\alpha(\tau)]\mathcal{D}[\mathbf{X}(\tau)], \tag{3.94}$$

in which $\mathcal{P}[\mathbf{X}(\tau),\alpha(\tau)]\mathcal{D}[\mathbf{X}(\tau)]$, is an infinite-dimensional probability distribution for the entire paths $[\mathbf{X}(\tau)]$.

Proof 5 (Jarzynski equality) *The basic idea for the derivation is as follows: We first represent a path $\mathbf{X}(t)$ by a discrete version with N steps and write the path probability in terms of the product of N transition probabilities given by the $\prod_{i=0}^{N}(\cdots)$ in Eq. (3.95). Then the mean-exponential of negative work*

$$\left\langle e^{-\frac{1}{\epsilon^2}W[\mathbf{X}(\tau),\alpha(\tau)]}\right\rangle_{\left[\mathbf{X}(\tau),\alpha(\tau)\right]}$$

is [64]:

$$\begin{aligned}&\left\langle e^{-\frac{1}{\epsilon^2}W[\mathbf{X}_0,\cdots,\mathbf{X}_N;\alpha_0,\cdots,\alpha_N]}\right\rangle_{\left[\mathbf{X}_0,\cdots,\mathbf{X}_N;\alpha_0,\cdots,\alpha_N\right]}\\ &=\int\cdots\int\prod_{i=0}^{N}d\mathbf{X}_i\exp\left(-\frac{1}{\epsilon^2}\sum_{i=1}^{N}\Big(\varphi(\mathbf{X}_i;\alpha_i)-\varphi(\mathbf{X}_i;\alpha_{i-1})\Big)\right)\\ &\quad\times\prod_{i=1}^{N}P\big(\mathbf{X}_i|\mathbf{X}_{i-1};\alpha_{i-1}\big)p\big(\mathbf{X}_0,t_0;\alpha_0\big),\end{aligned} \tag{3.95}$$

in which the work from state $(\mathbf{X}_i;\alpha_i)$ to state $(\mathbf{X}_{i+1};\alpha_{i+1})$ is defined as the difference in the global $\varphi(\mathbf{x};\alpha)$ with a common zero reference. This is a consequence of the First Law of Thermodynamics. Since equilibrium is attained at t_0, $p(\mathbf{x}_0;t_0)=f^{eq}_{\mathbf{X}}(\mathbf{x}_0;\alpha_0)$. With the global $\varphi(\mathbf{x};\alpha)=-\epsilon^2\ln f^{eq}_{\mathbf{X}}(\mathbf{x};\alpha)-\epsilon^2\ln Z(\alpha)$, we have:

$$\begin{aligned}&\left\langle e^{-\frac{1}{\epsilon^2}W[\mathbf{X}_0,\cdots,\mathbf{X}_N;\alpha_0,\cdots,\alpha_N]}\right\rangle_{\left[\mathbf{X}_N,\cdots,\mathbf{X}_0;\alpha_N,\cdots,\alpha_0\right]}\\ =&\int\cdots\int\prod_{i=0}^{N}d\mathbf{X}_i\prod_{i=1}^{N}\frac{f^{eq}_{\mathbf{X}}(\mathbf{X}_i;\alpha_i)Z(\alpha_i)}{f^{eq}_{\mathbf{X}}(\mathbf{X}_i;\alpha_{i-1})Z(\alpha_{i-1})}\cdot\prod_{i=1}^{N}P(\mathbf{X}_i|\mathbf{X}_{i-1};\alpha_{i-1})f^{eq}_{\mathbf{X}}(\mathbf{X}_0;\alpha_0)\\ =&\frac{Z(\alpha_n)}{Z(\alpha_0)}\int\cdots\int\prod_{i=0}^{N}d\mathbf{X}_i\prod_{i=1}^{N}P(\mathbf{X}_i|\mathbf{X}_{i-1};\alpha_{i-1})\prod_{i=1}^{N}f^{eq}_{\mathbf{X}}(\mathbf{X}_{i-1};\alpha_{i-1})\Bigg/\prod_{i=1}^{N}f^{eq}_{\mathbf{X}}(\mathbf{X}_i;\alpha_{i-1})\\ =&\frac{Z(\alpha_n)}{Z(\alpha_0)}.\end{aligned} \tag{3.96}$$

Since we have defined in Sec. 3.2.2 the free energy as: $A_2(\alpha)=-\epsilon^2\ln Z(\alpha)$, thus we obtain the Jarzynski equality:

$$\left\langle e^{-\frac{1}{\epsilon^2}W[\mathbf{X}(\tau),\alpha(\tau)]}\right\rangle_{\left[\mathbf{X}(\tau),\alpha(\tau)\right]}=e^{-\frac{1}{\epsilon^2}\Delta A_2}. \tag{3.97}$$

In a very similar vein, for the macroscopic thermodynamic variables $(E,\widetilde{\alpha})$, one defines the work done to the system by the external environment through controlling $\widetilde{\alpha}(t)$ with rate

$\dot{\widetilde{\alpha}}$:

$$W\big[E(\tau),\widetilde{\alpha}(\tau)\big] = -\int_0^t \widetilde{F}_{\widetilde{\alpha}}(E,\widetilde{\alpha})\dot{\widetilde{\alpha}}d\tau. \tag{3.98}$$

Write $\varphi_\tau(\widetilde{\alpha}) = E(\tau,\widetilde{\alpha})$. Then the discretized $\left\langle e^{-\frac{1}{\epsilon^2}W[E(\tau),\widetilde{\alpha}(\tau)]}\right\rangle_{\left[E(\tau),\widetilde{\alpha}(\tau)\right]}$ is:

$$\begin{aligned}
&\left\langle e^{-\frac{1}{\epsilon^2}W[E_0,\cdots,E_N;\widetilde{\alpha}_0,\cdots,\widetilde{\alpha}_N]}\right\rangle_{\left[E_0,\cdots,E_N;\widetilde{\alpha}_0,\cdots,\widetilde{\alpha}_N\right]}\\
=&\int\cdots\int\prod_{i=0}^N dE_i\prod_{i=1}^N P(E_i|E_{i-1};\widetilde{\alpha}_{i-1})p(E_0,t_0;\widetilde{\alpha}_0)\times\\
&\exp\left(-\sum_{i=1}^N\frac{\varphi_i(\widetilde{\alpha}_i)-\varphi_i(\widetilde{\alpha}_{i-1})}{\epsilon^2}+S(E_i,\widetilde{\alpha}_i)-S(E_i,\widetilde{\alpha}_{i-1})\right),
\end{aligned} \tag{3.99}$$

where $S(E,\widetilde{\alpha})$ is defined in Eq. (3.119). On the other hand, equilibrium probability density function of E at $(E_i,\widetilde{\alpha}_i)$ is:

$$\begin{aligned}
f_{\mathbf{E}}^{eq}(E_i,\widetilde{\alpha}_i) &= \frac{1}{Z(\widetilde{\alpha}_i)}\oint_{\varphi(\mathbf{x};\widetilde{\alpha})=E_i} e^{-\varphi(\mathbf{x};\widetilde{\alpha}_i)/\epsilon^2}\frac{d\Sigma^{n-1}}{||\nabla_{\mathbf{x}}\varphi(\mathbf{x};\widetilde{\alpha}_i)||}\\
&= Z^{-1}(\widetilde{\alpha}_i)\exp\left(-\frac{\varphi_i(\widetilde{\alpha}_i)}{\epsilon^2}+S(E_i,\widetilde{\alpha}_i)\right).
\end{aligned} \tag{3.100}$$

Hence, we have

$$\begin{aligned}
&\left\langle e^{-\frac{1}{\epsilon^2}W[E_0,\cdots,E_N;\widetilde{\alpha}_0,\cdots,\widetilde{\alpha}_N]}\right\rangle_{\left[E_0,\cdots,E_N;\widetilde{\alpha}_0,\cdots,\widetilde{\alpha}_N\right]}\\
=&\int\cdots\int\prod_{i=0}^N dE_i\ \prod_{i=1}^N\frac{f_{\mathbf{E}}^{eq}(E_i,\widetilde{\alpha}_i)Z(\widetilde{\alpha}_i)}{f_{\mathbf{E}}^{eq}(E_i,\widetilde{\alpha}_{i-1})Z(\widetilde{\alpha}_{i-1})}\left(\prod_{i=1}^N P(E_i|E_{i-1},\widetilde{\alpha}_{i-1})f_{\mathbf{E}}^{eq}(E_0,\widetilde{\alpha}_0)\right)\\
=&\frac{Z(\widetilde{\alpha}_n)}{Z(\widetilde{\alpha}_0)}\int\cdots\int\prod_{i=0}^N dE_i\ \prod_{i=1}^N P(E_i|E_{i-1},\widetilde{\alpha}_{i-1})\prod_{i=1}^N f_{\mathbf{E}}^{eq}(E_{i-1},\widetilde{\alpha}_{i-1})\Bigg/\prod_{i=1}^N f_{\mathbf{E}}^{eq}(E_i,\widetilde{\alpha}_{i-1})\\
=&\frac{Z(\widetilde{\alpha}_n)}{Z(\widetilde{\alpha}_0)} = e^{-\frac{1}{\epsilon^2}\Delta A_2(\alpha)}.
\end{aligned} \tag{3.101}$$

Therefore, the log-mean exponential of minus work is equal to the minus of free energy difference.

It is clear from the proof above that the Jarzynski equality is general for Markov processes with or without detailed balance.

Crooks' approach

G. E. Crooks' approach, when applied to processes without detailed balance [64], considers the probability functional of a backward path $\mathcal{P}[\check{\mathbf{X}}(t)|\check{\mathbf{X}}(0);\check{\alpha}(t)]$ over a forward one $\mathcal{P}[\mathbf{X}(t)|\mathbf{X}(0);\alpha(t)]$, where both the initial and final distribution of $\mathbf{X}(\tau)$ follows the equilibrium distribution:

$$\begin{aligned}\frac{\mathcal{P}[\check{\mathbf{X}}(\tau),\check{\alpha}(\tau)]}{\mathcal{P}[\mathbf{X}(\tau),\alpha(\tau)]} =& \exp\left(\frac{Q[\mathbf{X}(\tau),\alpha(\tau)]-Q_{hk}[\mathbf{X}(\tau),\alpha(\tau)]-\Delta\varphi+\Delta A_2}{\epsilon^2}\right)\\ =& \exp\left(\frac{-W[\mathbf{X}(\tau),\alpha(\tau)]-Q_{hk}[\mathbf{X}(\tau),\alpha(\tau)]+\Delta A_2}{\epsilon^2}\right), \qquad (3.102)\end{aligned}$$

where $\check{\mathbf{X}}(\tau)=\mathbf{X}(t-\tau)$, $\check{\alpha}(\tau)=\alpha(t-\tau)$; and

$$\begin{aligned} Q[\mathbf{X}(\tau),\alpha(\tau)] &= \int_0^t \frac{\partial\varphi}{\partial\mathbf{X}}\dot{\mathbf{X}}\,\mathrm{d}\tau,\\ Q_{hk}[\mathbf{X}(\tau),\alpha(\tau)] &= -2\int_0^t \dot{\mathbf{X}}^T D(\alpha)^{-1}\big(M(\alpha)-D(\alpha)\Xi^{-1}(\alpha)\big)\mathbf{X}\,\mathrm{d}\tau \qquad (3.103)\end{aligned}$$

are the heat dissipation and the house-keeping heat respectively.

If a process describes a physical system in equilibrium, which is expected to be "microscopic reversible" in [33], then

$$\left\langle\frac{\mathcal{P}[\check{\mathbf{X}}(\tau),\check{\alpha}(\tau)]}{\mathcal{P}[\mathbf{X}(\tau),\alpha(\tau)]}\right\rangle_{\left[\mathbf{X}(\tau),\alpha(\tau)\right]} = \int\mathcal{D}[\mathbf{X}(\tau),\alpha(\tau)]\;\mathcal{P}[\check{\mathbf{X}}(\tau),\check{\alpha}(\tau)] = 1. \qquad (3.104)$$

On the other hand, if the system is in detailed balance for each and every α, $M(\alpha)-D(\alpha)\Xi^{-1}(\alpha)=0$, then the house-keeping heat $Q_{hk}[\mathbf{X}(\tau),\alpha(\tau)]\equiv 0$. Therefore, path-ensemble average of Eq. (3.102) gives:

$$\begin{aligned} 1 &= e^{\Delta A_2/\epsilon^2}\left\langle e^{-W[\mathbf{X}(\tau),\alpha(\tau)]/\epsilon^2-Q_{hk}[\mathbf{X}(\tau),\alpha(\tau)]/\epsilon^2}\right\rangle_{\left[\mathbf{X}(\tau),\alpha(\tau)\right]}\\ &= e^{\Delta A_2/\epsilon^2}\left\langle e^{-W[\mathbf{X}(\tau),\alpha(\tau)]/\epsilon^2}\right\rangle_{\left[\mathbf{X}(\tau),\alpha(\tau)\right]}. \qquad (3.105)\end{aligned}$$

This is Hatano-Sasa's result [64]. For systems without detailed balance, $Q_{hk}[\mathbf{X}(\tau),\alpha(\tau)]$ measures the magnitude of the divergence-free vector field, or the extent to which the system is away from detailed balance, even when stationary distribution is attained. At the same time,

$$\left\langle\frac{\mathcal{P}[\check{\mathbf{X}}(\tau),\check{\alpha}(\tau)]}{\mathcal{P}[\mathbf{X}(\tau),\alpha(\tau)]}\right\rangle_{\left[\mathbf{X}(\tau),\alpha(\tau)\right]}$$

measures how much on average the behavior of backward paths is statistically different from forward ones.

Proof 6 (Crooks' approach) *Instead of introducing stochastic work functional, G. E. Crooks' approach recognizes the important role of* time reversal *trajectory* $\check{\mathbf{X}}(t)$*, and the deep relationship between work, energy, and dissipation, e.g., entropy production. Let us consider the path probability of a backward trajectory* $\mathcal{P}[\check{\mathbf{X}}(\tau)|\check{\mathbf{X}}(0);\check{\alpha}(\tau)]$ *against a forward one* $\mathcal{P}[\mathbf{X}(\tau)|\mathbf{X}(0);\alpha(\tau)]$*, in which* $\big(\check{\mathbf{X}}(\tau);\check{\alpha}(\tau)\big) = (\mathbf{X}(t-\tau);\alpha(t-\tau))$*, where the initial and final distribution of* $\mathbf{X}(\tau)$ *follow the equilibrium distribution. We solve for* $\mathcal{P}[\mathbf{X}(\tau)|\mathbf{X}(0);\alpha(\tau)]$ *from the probability of a Brownian motion whose increments are multivariate Gaussian:*

$$\frac{1}{\epsilon}\Gamma(\alpha_i)^{-1}\Big(\mathbf{X}_{i+1}-\mathbf{X}_i-M(\alpha_i)\mathbf{X}_i\Delta\tau\Big)=\mathbf{B}_{t_{i+1}}-\mathbf{B}_{t_i}. \tag{3.106}$$

Hence, the probability density functional of a path $\big[\mathbf{X}_0,\cdots,\mathbf{X}_N|\mathbf{X}_0;\alpha_0,\cdots,\alpha_N\big]$ *is:*

$$\begin{aligned}&\mathcal{P}\big[\mathbf{X}_0,\cdots,\mathbf{X}_N|\mathbf{X}_0;\alpha_0,\cdots,\alpha_N\big]=\prod_{i=0}^{N-1}P\big(\mathbf{X}_{i+1}|\mathbf{X}_i;\alpha_i\big)\\ =&\prod_{i=0}^{N-1}\frac{1}{(\pi\Delta\tau)^{n/2}}e^{-\frac{1}{\epsilon^2\Delta\tau}\big(\Gamma(\alpha_i)^{-1}(\mathbf{X}_{i+1}-\mathbf{X}_i)-\Gamma(\alpha_i)^{-1}M(\alpha_i)\mathbf{X}_i\Delta\tau\big)^2}.\end{aligned} \tag{3.107}$$

Here $\big(\mathbf{v}(\mathbf{X},\alpha)\big)^2\equiv\big(\mathbf{v}(\mathbf{X},\alpha)\big)^T\big(\mathbf{v}(\mathbf{X},\alpha)\big)$*. Then the probability density functional of the inverse path* $\big[\mathbf{X}_N,\cdots,\mathbf{X}_0|\mathbf{X}_N;\alpha_N,\cdots,\alpha_0\big]$ *is:*

$$\begin{aligned}&\mathcal{P}\big[\mathbf{X}_N,\cdots,\mathbf{X}_0|\mathbf{X}_N;\alpha_N,\cdots,\alpha_0\big]=\prod_{i=1}^{N}P\big(\mathbf{X}_{i-1}|\mathbf{X}_i;\alpha_i\big)\\ =&\prod_{i=1}^{N}\frac{1}{(\pi\Delta\tau)^{n/2}}e^{-\frac{1}{\epsilon^2\Delta\tau}\big(\Gamma(\alpha_i)^{-1}(\mathbf{X}_{i-1}-\mathbf{X}_i)-\Gamma(\alpha_i)^{-1}M(\alpha_i)\mathbf{X}_i\Delta\tau\big)^2}.\end{aligned} \tag{3.108}$$

Therefore, offsetting by a normalization factor, an infinite-dimensional functional integral,

$$\begin{aligned}\mathcal{P}\big[\mathbf{X}(\tau)|\mathbf{X}(0);\alpha(\tau)\big]&\propto\exp\left[-\frac{1}{\epsilon^2}\int_0^t\Big(\Gamma(\alpha)^{-1}d\mathbf{X}/\sqrt{d\tau}-\Gamma(\alpha)^{-1}M(\alpha)\mathbf{X}\sqrt{d\tau}\Big)^2\right]\\ &=\exp\left[-\frac{1}{\epsilon^2}\int_0^t\Big(\Gamma(\alpha)^{-1}\dot{\mathbf{X}}-\Gamma(\alpha)^{-1}M(\alpha)\mathbf{X}\Big)^2d\tau\right].\end{aligned} \tag{3.109}$$

Probability of the backward path can be found by substituting τ *with* $t-\tau$*:*

$$\mathcal{P}\big[\check{\mathbf{X}}(\tau)|\check{\mathbf{X}}(0);\check{\alpha}(\tau)\big]\propto\exp\left[-\frac{1}{\epsilon^2}\int_0^t\Big(-\Gamma(\alpha)^{-1}\dot{\mathbf{X}}-\Gamma(\alpha)^{-1}M(\alpha)\mathbf{X}\Big)^2d\tau\right]. \tag{3.110}$$

Therefore, we have an equality for heat dissipation:

$$
\begin{aligned}
&\frac{\mathcal{P}[\check{\mathbf{X}}(\tau)|\check{\mathbf{X}}(0);\check{\alpha}(\tau)]}{\mathcal{P}[\mathbf{X}(\tau)|\mathbf{X}(0);\alpha(\tau)]}\\
&= \exp\left[\frac{2}{\epsilon^2}\int_0^t \left(\dot{\mathbf{X}}^T(\Gamma(\alpha)\Gamma(\alpha)^T)^{-1}M(\alpha)\mathbf{X} + \mathbf{X}^T M(\alpha)^T(\Gamma(\alpha)\Gamma(\alpha)^T)^{-1}\dot{\mathbf{X}}\right)\, d\tau\right]\\
&= \exp\left[\frac{2}{\epsilon^2}\int_0^t \left(\dot{\mathbf{X}}^T D^{-1}(\alpha)M(\alpha)\mathbf{X}\right)\, d\tau\right]\\
&= \exp\left[\frac{2}{\epsilon^2}\int_0^t \dot{\mathbf{X}}^T \Xi^{-1}(\alpha)\mathbf{X}\, d\tau + \frac{2}{\epsilon^2}\int_0^t \dot{\mathbf{X}}^T\big(D^{-1}(\alpha)M(\alpha) - \Xi^{-1}(\alpha)\big)\mathbf{X}\, d\tau\right]\\
&= \exp\left\{\frac{Q\big[\mathbf{X}(\tau),\alpha(\tau)\big]}{\epsilon^2} + \frac{Q_{hk}\big[\mathbf{X}(\tau),\alpha(\tau)\big]}{\epsilon^2}\right\}. \qquad (3.111)
\end{aligned}
$$

Hatano and Sasa, following Oono and Paniconi, called the term

$$
Q_{hk}\big[\mathbf{X}(\tau),\alpha(\tau)\big] = -2\int_0^t \dot{\mathbf{X}}^T\big(D^{-1}(\alpha)M(\alpha) - \Xi^{-1}(\alpha)\big)\mathbf{X} d\tau \qquad (3.112)
$$

house-keeping heat [64].

Since we start and end with equilibrium distributions with the corresponding $\widetilde{\alpha}$*,*

$$
\begin{aligned}
p\big(\mathbf{X}_0;\alpha_0\big) &= f^{eq}_{\mathbf{X}}(\mathbf{X}_0;\alpha_0) = \frac{1}{Z(\alpha_0)}\exp\left[-\frac{\mathbf{X}_0^T U(\alpha_0)\mathbf{X}_0}{\epsilon^2}\right];\\
p\big(\mathbf{X}_N;\alpha_N\big) &= f^{eq}_{\mathbf{X}}(\mathbf{X}_N;\alpha_N) = \frac{1}{Z(\alpha_N)}\exp\left[-\frac{\mathbf{X}_N^T U(\alpha_N)\mathbf{X}_N}{\epsilon^2}\right]. \qquad (3.113)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\frac{\mathcal{P}[\check{\mathbf{X}}(\tau),\check{\alpha}(\tau)]}{\mathcal{P}[\mathbf{X}(\tau),\alpha(\tau)]} &= \frac{\mathcal{P}[\check{\mathbf{X}}(\tau)|\check{\mathbf{X}}(0);\check{\alpha}(\tau)]p(\mathbf{X}_N;\alpha_N)}{\mathcal{P}[\mathbf{X}(\tau)|\mathbf{X}(0);\alpha(\tau)]p(\mathbf{X}_0;\alpha_0)}\\
&= \exp\left(\frac{Q[\mathbf{X}(\tau),\alpha(\tau)] - Q_{hk}[\mathbf{X}(\tau),\alpha(\tau)] - \Delta\varphi + \Delta A_2}{\epsilon^2}\right)\\
&= \exp\left(\frac{-W[\mathbf{X}(\tau),\alpha(\tau)] - Q_{hk}[\mathbf{X}(\tau),\alpha(\tau)] + \Delta A_2}{\epsilon^2}\right). \qquad (3.114)
\end{aligned}
$$

Now taking ensemble average of the trajectories $[\mathbf{X}(\tau),\alpha(\tau)]$ *over* $\dfrac{\mathcal{P}[\check{\mathbf{X}}(\tau),\check{\alpha}(\tau)]}{\mathcal{P}[\mathbf{X}(\tau),\alpha(\tau)]}$ *gives:*

$$
\begin{aligned}
\int \mathcal{D}[\mathbf{X}(\tau),\alpha(\tau)]\,\mathcal{P}[\check{\mathbf{X}}(\tau),\check{\alpha}(\tau)] &= \left\langle\frac{\mathcal{P}[\check{\mathbf{X}}(\tau),\check{\alpha}(\tau)]}{\mathcal{P}[\mathbf{X}(\tau),\alpha(\tau)]}\right\rangle_{[\mathbf{X}(\tau),\alpha(\tau)]}\\
= \quad e^{\Delta A_2}\left\langle e^{-W[\mathbf{X}(\tau),\alpha(\tau)] - Q_{hk}[\mathbf{X}(\tau),\alpha(\tau)]}\right\rangle_{\big[\mathbf{X}(\tau),\alpha(\tau)\big]}&. \qquad (3.115)
\end{aligned}
$$

When the system is in detailed balance, Crooks' approach recovers the Jarzynski equality. If one chooses the global energy φ with zero reference for each own equilibrium, i.e., $\Delta A_2 = 0$ for all α, then it recovers the Hatano-Sasa equality.

Crooks' approach through adjoint processes Jarzynski's approach is based on a mesoscopic α-force; while Crooks' approach concerns with the stochastic entropy production rate which reflects "heat dissipation". Therefore, for systems with detailed balance, they are essentially the same result according to the First Law of thermodynamics. For systems without detailed balance, one can again obtained a Jarzynski-like equality from the probability P of the forward path over the adjoint probability $P^\dagger$ of the backward one, according to the notion of *time reversal* in Eq. (3.60b):

$$\begin{aligned} P^\dagger(\mathbf{X}(\tau)|\mathbf{X}(\tau+\mathrm{d}\tau);\alpha(\tau)) =& P(\mathbf{X}(\tau+\mathrm{d}\tau)|\mathbf{X}(\tau);\alpha(\tau)) \\ & \times \left(\frac{f^{eq}_{\mathbf{X}}(\mathbf{X}(\tau);\alpha(\tau))}{f^{eq}_{\mathbf{X}}(\mathbf{X}(\tau+\mathrm{d}\tau);\alpha(\tau))} \right). \end{aligned} \tag{3.116}$$

Thus, whether a system is in detailed balance or not, one has the Hatano-Sasa equality [64]:

$$1 = \left\langle \frac{\mathcal{P}[\mathbf{X}(\tau),\alpha(\tau)]}{\mathcal{P}^\dagger[\check{\mathbf{X}}(\tau),\check{\alpha}(\tau)]} \right\rangle_{\left[\mathbf{X}(\tau),\alpha(\tau)\right]} = e^{\Delta A_2/\epsilon^2} \left\langle e^{-W[\mathbf{X}(\tau),\alpha(\tau)]/\epsilon^2} \right\rangle_{\left[\mathbf{X}(\tau),\alpha(\tau)\right]}. \tag{3.117}$$

Macroscopic work equalities

We are now in the position to study the work-free energy relation from a macroscopic view. Essentially, we will consider the stochastic, fluctuating $(E(t),\widetilde{\alpha}(t))$ instead of $(\mathbf{X}(t);\alpha(t))$ directly. In doing so, we are observing the evolution in the probability distribution of E through a projection from $(\mathbf{X};\alpha)$ to $(E,\widetilde{\alpha})$. With the projection of the n-dimensional phase space to the one-dimensional time series $E(t)$, the stationary probability density function $f^{ss}_{\mathbf{E}}(E,\widetilde{\alpha})$ of E with $\widetilde{\alpha}$ is also a projection of the original stationary probability density function $f^{ss}_{\mathbf{X}}(\mathbf{x};\widetilde{\alpha})$ in Euclidean space (as discussed in Sec. 3.2.3):

$$\begin{aligned} f^{ss}_{\mathbf{E}}(E,\widetilde{\alpha}) &= \oint_{\varphi(\mathbf{x};\widetilde{\alpha})=E} \frac{f^{ss}_{\mathbf{X}}(\mathbf{x};\widetilde{\alpha})\,\mathrm{d}\Sigma^{n-1}}{||\nabla_{\mathbf{x}}\varphi(\mathbf{x};\widetilde{\alpha})||} \\ &= \frac{1}{Z(\widetilde{\alpha})}\exp\left(-\frac{E}{\epsilon^2}+S(E,\widetilde{\alpha})\right), \end{aligned} \tag{3.118}$$

in which

$$S(E,\widetilde{\alpha}) = \ln\left(\oint_{\varphi(\mathbf{x};\widetilde{\alpha})=E} \frac{\mathrm{d}\Sigma^{n-1}}{||\nabla_{\mathbf{x}}\varphi(\mathbf{x};\widetilde{\alpha})||} \right). \tag{3.119}$$

For the process of $(E(t), \widetilde{\alpha}(t))$, the total internal energy is no longer E itself. But rather, it would include the "entropic effect", $S(E, \widetilde{\alpha})$, caused by the curved space structure, and become $\mathscr{A}(E, \widetilde{\alpha})$.

$$\mathscr{A}(E, \widetilde{\alpha}) = E - \epsilon^2 S(E, \widetilde{\alpha}). \tag{3.120}$$

The Helmholtz theorem for the new $(E(t), \widetilde{\alpha}(t))$ process reads:

$$\begin{aligned} \mathrm{d}\mathscr{A} &= \widetilde{\Theta}(E, \widetilde{\alpha})\mathrm{d}\sigma - \widetilde{F}_{\widetilde{\alpha}}(E, \widetilde{\alpha})\mathrm{d}\widetilde{\alpha} \\ &= \left(\left(\frac{\partial \sigma}{\partial E} \right)^{-1} - \epsilon^2 \frac{\partial S}{\partial \sigma} \right) \mathrm{d}\sigma - \left(\left(\frac{\partial \sigma}{\partial \widetilde{\alpha}} \right) \left(\frac{\partial \sigma}{\partial E} \right)^{-1} + \epsilon^2 \frac{\partial S}{\partial \widetilde{\alpha}} \right) \mathrm{d}\widetilde{\alpha}. \end{aligned} \tag{3.121}$$

Hence, the total force that does the work in this new coordinate is:

$$\widetilde{F}_{\widetilde{\alpha}}(E, \widetilde{\alpha}) = -\frac{\partial \mathscr{A}(E, \widetilde{\alpha})}{\partial \widetilde{\alpha}} = \frac{1}{n} \cdot \frac{E}{\widetilde{\alpha}} + \epsilon^2 \frac{\partial S(E, \widetilde{\alpha})}{\partial \widetilde{\alpha}}, \tag{3.122}$$

where $\epsilon^2 \big(\partial S(E, \widetilde{\alpha})/\partial \widetilde{\alpha}\big)$ is what chemists called an "entropic force".

Now we define the work that external environment has done to the system through the controlled change of $\widetilde{\alpha}(t)$ as:

$$W[E(\tau), \widetilde{\alpha}(\tau)] = -\int_0^t \widetilde{F}_{\widetilde{\alpha}}(E, \widetilde{\alpha})\dot{\widetilde{\alpha}}\mathrm{d}\tau. \tag{3.123}$$

Then the work-free energy relation in macroscopic variables is:

$$\left\langle e^{-W[E(\tau), \widetilde{\alpha}(\tau)]} \right\rangle_{\left[E(\tau), \widetilde{\alpha}(\tau)\right]} = \frac{Z\big(\widetilde{\alpha}(t)\big)}{Z\big(\widetilde{\alpha}(0)\big)} = \exp\left(-\frac{\Delta A_2(\alpha)}{\epsilon^2} \right). \tag{3.124}$$

Therefore, the averaged minus exponential of work is equal to the minus exponential of free energy difference. Here, we notice that the free energy stays the same through the change of free variables, as a result of Eq. (3.86):

$$Z(\alpha) = \int_{\mathbb{R}^n} e^{-\varphi(\mathbf{x};\alpha)/\epsilon^2} \mathrm{d}\mathbf{x} = \int_{\varphi_{min}(\alpha)}^{\infty} \exp\left(-\frac{E}{\epsilon^2} + S(E, \alpha) \right) \mathrm{d}E.$$

3.2.5 Discussion

In the present work, using the OUP as an example, we have illustrated a possible method of deriving emergent, macroscopic descriptions of a complex stochastic dynamics from its mesoscopic *law of motion*. In recent years, there is a growing awareness of the role of probabilistic reasoning as *the* logic of science [75, 129]. In this framework, prior information, data, and probabilistic deduction are three pillars of a scientific theory. In fields with very complex dynamics, statistical inferences focus on the latter two aspects starting with data. In physical sciences that includes chemistry, and cellular biology, the prior plays a fundamental role as a feasible "mechanism" which enters a scientific model based on "established knowledge" — no biochemical phenomena should violate the physical laws of mechanics and thermodynamics. Indeed, many priors have been rigorously formalized in terms of mathematical theories. Unfortunately, most of these theories are expressed in terms of deterministic mathematics for very simple individual "particles"; obtaining a meaningful probabilistic prior for a realistic, macroscopic-level system requires a computational task that is neither feasible nor meaningful [6, 141]. Nonlinear stochastic dynamical study is the mathematical deductive process that formulates probabilistic prior based on a given mechanism.

Open systems, when represented in terms of Markov processes, are ubiquitously non-symmetric processes according to Kolmogorov's terminology. This is one of the lessons we learned from the open-chemical systems theory. The non-symmetricity can be quantified by *entropy production* [76]. For discrete-state Markov processes, symmetric processes are equivalent to Kolmogorov's cycle condition [78]. Interestingly, concepts such as cycle condition, detailed balance, dissipation and irreversible entropy production had all been independently discovered in chemistry: Wegscheider's relation in 1901 [181], detailed balance by G.N. Lewis in 1925 [97], Onsager's dissipation function in 1931 [115], and the formulation of entropy production in the 1940s [130, 166].

A non-symmetric Markov process implies circulating dynamics in phase space. Such dynamics is not necessarily dissipative, as exemplified by harmonic oscillators in classical mechanics. One of us has recently pointed out the important distinction between *overdamped thermodynamics* and *underdamped thermodynamics* [**?**]. The current section is a study of OUP in terms of the latter perspective, in which we have identified the unbalanced circulation as a conservative dynamics, a hallmark of the generalized underdamped thermodynamics [136]. In terms of this conservative dynamics, Boltzmann's entropy function naturally enters stochastic thermodynamics, and we discover a relation between the Helmholtz theorem [104] and the various work relations.

In the past, studies on stochastic thermodynamics with underdamped mechanical motions have always required an explicit identification of even and odd variables. See [**?**] and the references cited within. One of us has introduced a more general stochastic formulation of "underdamped" dynamics, with thermodynamics, in which circulating motion can be a part of a conservative motion [**?**] without dissipation. The present work is an in-depth study of the OUP within this new framework. It seems to us that even the term "nonequilibrium" in the literature has two rather different meanings: From a classical mechanical standpoint, any system with a stationary current is "nonequilibrium", even though it can be non-dissipative. From a statistical mechanics stand point, on the other hand, "nonequilibrium", "irreversible", and "dissipative" are almost all synonymous.

Absolute information theory and interpretive information theories

We now discuss two rather different perspectives on the nature of information theory, or theories [80, 75].

First, in the framework of classical physics in terms of Newtonian mechanics, Boltzmann's law, and Gibbs' theory of chemical potential, there is a universal First Law of Thermodynamics based on the function $S(E, V, N, \alpha)$ where S is the Boltzmann's entropy of a conservative dynamical system at total energy E, e.g., Hamiltonian $H\big(\{x_i\}, \{y_i\}\big) = E$, with V and $N = (n_1, n_2, \cdots, n_m)$ being the volume and numbers of particles in the chemo-mechanical system, and $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_\nu)$ represents controllable parameters of the system. Then one has

$$\mathrm{d}E = \left(\frac{\partial S}{\partial E}\right)^{-1}_{V,N,\alpha} \mathrm{d}S - p\mathrm{d}V + \mu\mathrm{d}N - \left(\frac{\partial S}{\partial E}\right)^{-1}_{V,N,\alpha}\left(\frac{\partial S}{\partial \alpha}\right)_{E,V,N} \mathrm{d}\alpha, \qquad (3.125)$$

in which $(\partial S/\partial E)^{-1}_{V,N,\alpha}$ is absolute temperature. p and μ are pressure and chemical potential, they are the corresponding thermodynamic forces for changing volume V and number of particles N, respectively. It is natural to suggest that if an agent is able to manipulate a classical system through changing α while holding S, V, and N constant, then he or she is providing to, or extracting from, the classical system *non-mechanical, non-chemical work*. It will be the origin of a Maxwell's demon [106].

For an isothermal system, one can introduce Helmholtz's free energy function $A = E - TS$, then Eq. (3.125) becomes

$$\mathrm{d}A = \mu\mathrm{d}N - S\mathrm{d}T - p\mathrm{d}V + F_\alpha \mathrm{d}\alpha. \qquad (3.126)$$

And for an isothermal, isobaric information manipulation process without chemical reactions, one has Gibbs function $G = E - TS + pV$ and $\mathrm{d}G = \mu\mathrm{d}N - S\mathrm{d}T + V\mathrm{d}p + F_\alpha\mathrm{d}\alpha = F_\alpha\mathrm{d}\alpha$. Note that while the first three terms contain "extensive" quantities N, S, and V, the last term usually does not. It is nanothermodynamic [65]. Note also that for a feedback system that controls F_α, one has $\Theta = G - F_\alpha\alpha$ and $\mathrm{d}\Theta = -\alpha\mathrm{d}F_\alpha$.

Just as μ is a function of temperature T in general, so is F_α: It has an entropic part [138, 132]. This is where the "information" in Maxwell's demon enters thermodynamics.

Eqs. (3.125) and (3.126), thus, are a grander First Law which now includes feedback information as a part of the conservation [168] with "informatic energy" $F_\alpha \mathrm{d}\alpha$, on a par with heat energy $T\mathrm{d}S$, mechanical energy $p\mathrm{d}V$, and Gibbs' chemical energy $\mu\mathrm{d}N$. Eq. (3.125) is the theory of absolute information in connection to controlling α.

In engineering and biological research on complex systems, however, the notion of information often has a more subjective meaning, or meanings, usually hidden in the form of a statistical prior [128, 127]. One of the best examples, perhaps, is in current cellular biology: Many key biochemical processes inside a living cell are said to be "carrying out cellular signal transduction". Various biochemical activities and changing molecular concentrations are "interpreted" as "intracellular signals" that instruct a cell to respond to its environment. Here, two very different, but complementary, mathematical theories are equally valid: Since nearly all cellular biochemical reactions can be considered at constant temperature and volume, one describes the stochastic biochemical dynamics in terms of Gibbs' theory based on the μ in Eq. (3.126). On the other hand, the same stochastic biochemical dynamics described in term of the probability theory can also be represented as an information processing machine with communication channels and transmissions of bits of information, carrying out a myriad of biological functions such as sensing, proofreading, timing, adaptation, and amplifications of signal magnitude, detection sensitivity, and response specificity [62]. The information flow narratives provide bioscientists a higher level of abstraction of a physicochemical reality [140].

Such an interpretive information theory, however, will lack the fundamental character of Eqs. (3.125) and (3.126). Still, as a multi-scale, coarse grained theory, some inequalities can be established [154, 41]. It is also noted that changing α can always be mechanistically further represented in terms of changing geometric quantities such as volume and particle numbers via chemical reactions: The ultimate physical bases of information and its manipulation have to be matters and known forces.

We believe this dual possibility has a fundamental reason, rooted in Kolmogorov's rigorous theory of probability: A probability space is an abstract object associated with which many different random variables, as measurements, are possible. At this point, it is interesting to read the preface of [75] written by E. T. Jaynes, who is considered by many as one of the greatest information theorists since Shannon: "From many years of experience with its applications in hundreds of real problems, our views on the foundations of probability theory have evolved into something quite complex, which cannot be described in any such simplistic terms as 'pro-this' or 'anti-that'. For example, our system of probability could hardly be more different from that of Kolmogorov, in style, philosophy, and purpose. What we consider to be fully half of probability theory as it is needed in current applications —

the principles for assigning probabilities by logical analysis of incomplete information — is not present at all in the Kolmogorov system."

Then in an amazing candidness, Jaynes goes on: "Yet, when all is said and done, we find ourselves, to our own surprise, in agreement with Kolmogorov and in disagreement with its critics, on nearly all technical issues."

Chapter 4

ANALYZING NOISE INDUCED PHENOMENA THROUGH REVERSIBILITY

The decomposition framework in Sec. 2.3 of the general Markov processes with respect to reversibility can have a wide range of applications. Those applications can be oriented towards statistical methods, such as proposing Markov processes with the desired stationary distributions to generate samples in Monte Carlo algorithms (as will be discussed in Chapters 5 and 6). They can also be directed towards modelling. In particular, the decomposition can be used to analyze the dynamics of models when noise is taken into account. It builds up a bridge between the noisy system and its most probable dynamics in the sense that the most probable dynamics follow the landscape of the stationary distribution. Hence we can analyze the stability and bifurcation behaviors of this most probable effective dynamics to understand the stochastic system.

In this chapter, we will focus on this perspective. We are especially interested in the change of dynamical behaviors in a reaction diffusion system after noise is taken into account in the model. In particular, we study the stochastic reaction diffusion models where the original deterministic model is the infinite population limit of it. The inclusion of chemical reaction noise (with finite population size) induces pattern formation not present in the deterministic (inifite population) counterpart. Since one would usually assume the role of fluctuation and noise to be destructive to the organized patterns, the phenomenon of chemical reaction niose to play an organizing role has attracted much attention. Initial efforts are made to discover and categorize the observations. Later, some of the works start to use renormalization group theory to analyze the models one by one. With our framework in Sec. 2.3, however, we can start analyzing the effective dynamics (its stability and bifurcation) of the noisy system following its stationary distribution. In Sec. 4.1, we use this idea to analyze the Gray-Scott model and found that taking chemical reaction noise into account does cause the effective dynamics to become less stable and eventually Turing unstable.

4.1 Chemical Reaction Noise Induced Phenomena in Gray Scott Model: Change in Dynamics and Pattern Formation

Recent success of reaction diffusion models in digit patterning [144], oogenesis [118, 163], and etc. has aroused much interest on patterns generated by simple reaction diffusion systems [171, 167]. Much of the modeling effort has been done using deterministic systems. Those models describes the scenarios where size of the whole system takes infinite limit. In reality, however, the system size is often finite and stochasticity need to be taken into account. Specifically, behaviors of most biological and chemical systems are affected by the intrinsic chemical reaction noise. It has also been previously observed that noise has an important role in shaping the patterns formed in the reaction diffusion systems [95, 108].

Some previous works have simulated reaction diffusion systems with and without noise through numerically experiments, and have observed different behaviors. But the exact role of noise and its mechanism of changing the systems' behaviors remains unclear as the powerful tools like stability and bifurcation analysis are not present in the stochastic systems. One of the obstacles for applying these quantitative and qualitative tools in stochastic dynamics is that it's not previously known exactly what's the "effective" dynamics that caused the systems' behavioral difference, since every sample of the solutions is different.

Since the formation of complex patterns discussed are caused by the systems' stationary behaviors, the effective dynamics should correspond to the equilibrium distribution of the stochastic dynamics. In this work, we adopt the framework in [103, 102] and Sec. 2.3 to calculate effective dynamics for a quantitative analysis on pattern formation induced by chemical reaction noise. The effective dynamics is naturally consistent with the systems' long term distribution.

We focus on the Gray-Scott model [124], a classical model with rich behaviors of pattern formation. We derived the form of fluctuation from the reaction part of the model. It is found that under some parameter choice, the system behaves vastly different between cases where noise is present or not. The first example in this paper shows the organizing role of the noise, that pattern begins to form in the presence of noise. The second example shows that the pattern changes when noise is introduced. Analysis on the effective dynamics reveals the changes in stability that leads to the differences in complex patterns.

4.1.1 Gray-Scott Model with Chemical Reaction Noise

The Gray-Scott model describes a nonlinear chemical reaction system with two reactions and two components, U and V:

$$A + 2B \xrightarrow{1} 3B, \;\; B \xrightarrow{k} P. \tag{4.1}$$

Without losee of generality, one can assume that the rate constant for the first reaction to be unity.

There are currently two main approches to study the *macroscopic limit* of a stochastic reaction diffusion system, such as that in (4.1), with the molecular species also undergo spatial diffusion. One follows the celebrated work of Guo, Papanicolaou, and Varadhan (GPV) in terms of the empirical measure of the stochastic system [60, 44], and another [155].

We consider the following setup: A three dimensional system has a spatical inhomogeneity in the xy two-dimension but rapidly stirred in the x dimension. The xy plane is discretized into lattice sites $\{\mathbf{x_i}\}$ with l^2-sized "voxes", and associated with each vox a column with height H. We use $\widehat{U}(\mathbf{x_i})$ and $\widehat{V}(\mathbf{x_i})$ to represent the numbers of A and B in the column located at $\mathbf{x_i} \in \mathbb{R}^2$. Then the numbers of A and B per unit area are $U(\mathbf{x_i}) = \widehat{U}(\mathbf{x_i})/l^2$ and $V(\mathbf{x_i}) = \widehat{V}(\mathbf{x_i})/l^2$, and the three dimensional number density at $\mathbf{x_i}$ is $u(\mathbf{x_i}) = U(\mathbf{x_i})/H$ and $v(\mathbf{x_i}) = V(\mathbf{x_i})/H$.

Thermodynamic limit is defined as $U(\mathbf{x}_i), V(\mathbf{x}_i), H \to \infty$ while $u(\mathbf{x_i}) = U(\mathbf{x_i})/H$ and $v(\mathbf{x_i}) = H(\mathbf{x_i})/H$ becomes continous functions on 2-dimensional discrete lattice. *Hydrodynamic limit*, on the other hand, is defined as the $\ell \to 0$ and $u(\mathbf{x_i})$ and $v(\mathbf{x_i})$ become continuous functions $u(\mathbf{x})$ and $v(\mathbf{x})$ of continuous space variable $\mathbf{x} \in \mathbb{R}^2$.

Deterministic limit of the reaction dynamics. First, let us assume the molecules in a single vox column are rapidly mixed but do not move into the neighboring columns. Then in the limit of $H \to \infty$, according to T. G. Kurtz's theorem, the law of mass action arises with a set of differential equations at each and every lattice point $\mathbf{x_i}$:

$$\begin{cases} \mathrm{d}u = -uv^2 \mathrm{d}t, \\ \mathrm{d}v = (uv^2 - kv)\mathrm{d}t. \end{cases} \tag{4.2}$$

Stochastic chemical reaction. For a finite H, a chemical master equation characterizes the probability distribution of the finite population of chemical species within each colume

$$\begin{aligned}\frac{\mathrm{d}P(\mu,\nu)}{\mathrm{d}t} &= -P(\mu,\nu)\left(\frac{\mu\nu^2}{H^2}+k\nu\right)+P(\mu,\nu+1)k(\nu+1)\\ &\quad +P(\mu+1,\nu-1)\frac{(\mu+1)(\nu-1)^2}{H^2},\end{aligned} \tag{4.3}$$

in which $P(\mu,\nu) = \Pr\{U=\mu, V=\nu\}$ is the probability of observing μ number of U molecules and ν number of V molecules in a vox column of height H. We use Kramers Moyal expansion [104, 86] of Eq. (4.3) with respect to $\theta = 1/\Omega$ and write $p(u,v) = P(U,V)$. Through Itô's convention, the Fokker-Planck equation for the reaction system can be written as [3, 59]:

$$\frac{\partial p}{\partial t} =\theta\nabla^2 : \big(\mathbf{D}(u,v)p\big) - \nabla^T\cdot\big(\mathbf{F}_r(u,v)p\big), \tag{4.4}$$

where

$$\mathbf{D}(u,v) = \frac{1}{2}\begin{pmatrix} uv^2 & -uv^2 \\ -uv^2 & uv^2+kv \end{pmatrix}; \tag{4.5}$$

$$\mathbf{F}_r(u,v) = \begin{pmatrix} -uv^2 \\ uv^2-kv \end{pmatrix}. \tag{4.6}$$

Under Itô's interpretation, the Fokker-Planck equation represents the evolution of probability density function of the following stochastic differential equation (SDE):

$$\begin{pmatrix} \mathrm{d}u \\ \mathrm{d}v \end{pmatrix} = \mathbf{F}_r(u,v)\mathrm{d}t + \theta^{\frac{1}{2}}\sigma(u,v)\mathrm{d}\mathbf{W}(t), \tag{4.7}$$

where

$$\sigma(u,v) = \begin{pmatrix} \frac{\mu(\eta-k)+\nu(k+\eta)}{2\sqrt{2}\eta} & \frac{uv(\nu-\mu)}{\sqrt{2}\eta} \\ \frac{uv(\nu-\mu)}{\sqrt{2}\eta} & \frac{\nu(\eta-k)+(k+\eta)\mu}{2\sqrt{2}\eta} \end{pmatrix}, \quad \sigma(u,v)\sigma^T(u,v) = 2\mathbf{D}(u,v), \tag{4.8}$$

with $\eta = \sqrt{k^2+4u^2v^2}$, $\mu = \sqrt{v\,(k+2uv+\eta)}$, $\nu = \sqrt{v\,(k+2uv-\eta)}$.

It is worth noting that when we take $\theta \to 0$, Eq. reduces to Eq. (law of mass action).

Spatial diffusion and environmental input Apart from the chemical reaction at each site, the Gray-Scott model also include two terms: one diffusive term, $\mathbf{M}(u, v)$, over physical space $\mathbf{x}$, describing diffusion of chemical components through the space over reaction sites; and an deterministic input term, $\mathbf{F}_n(u, v)$, accounting for the feed for U and drain for both U and V. Hence, the complete SDE in the physical space $\mathbf{x}$ is:

$$\begin{pmatrix} \dfrac{\partial u(\mathbf{x}, t)}{\partial t} \\ \dfrac{\partial v(\mathbf{x}, t)}{\partial t} \end{pmatrix} = \mathbf{M}(u, v) + \mathbf{F}(u, v) + \theta^{\frac{1}{2}} \sigma(u, v) \xi(\mathbf{x}, t), \tag{4.9}$$

where

$$\mathbf{M}(u, v) = \begin{pmatrix} M_u \triangle_{\mathbf{x}} u \\ M_v \triangle_{\mathbf{x}} v \end{pmatrix}, \tag{4.10}$$

$$\mathbf{F}(u, v) = \mathbf{F}_r(u, v) + \mathbf{F}_n(u, v) = \begin{pmatrix} -uv^2 + f(1-u) \\ uv^2 - (f+k)v \end{pmatrix}, \tag{4.11}$$

$$\left\langle \xi(\mathbf{x}, t) \xi^T(\mathbf{x}', t') \right\rangle = \delta(\mathbf{x} - \mathbf{x}') \delta(t - t') \mathrm{I}. \tag{4.12}$$

In this work, we consider slow diffusion relative to the speed of reaction. We take $M_u = 4 \times 10^{-5}$, and $M_v = 2 \times 10^{-5}$. And for this reason, we solely focus on the intrinsic noise generated by the reaction part of the system, without taking into account the space-correlated noise caused by the diffusion effect.

4.1.2 Noise Induced Phenomena in Pattern Formation

It has been observed that noise can affect pattern formation. We hereby demonstrate and study two examples that noise can induce or change patterns. In the first example, we take $f = 0.053$, $k = 0.06$. When noise is not present, no spatial pattern is generated by the system; and when noise is present (by the intensity of $\theta = 2 \times 10^{-3}$), the dot-like patterns are generated. In the second example, we choose $f = 0.0483$, $k = 0.06$, so that the dot-like patterns are generated when noise is not present. Then we find that the dot-like patterns change to strip-like when noise is introduced to the system with an intensity of $\theta = 3 \times 10^{-3}$.

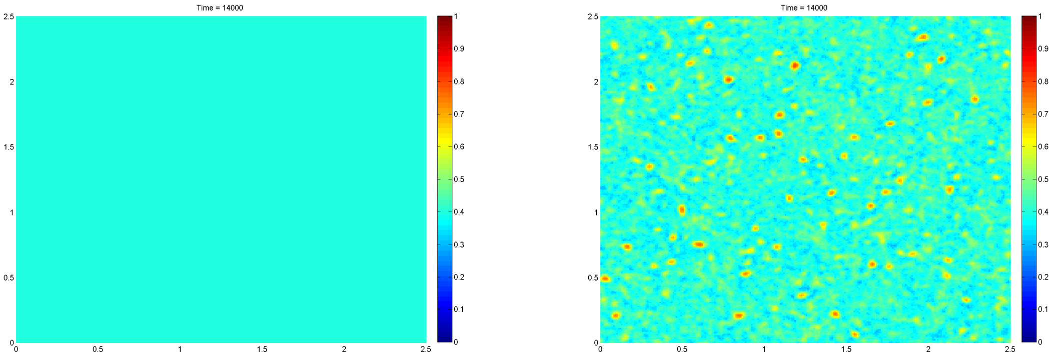


Figure 4.1: Left panel. Long time behavior of the Gray-Scott model without noise ($f = 0.053$, $k = 0.06$, $M_u = 4 \times 10^{-5}$, $M_v = 2 \times 10^{-5}$); Right panel. Long time behavior of the Gray-Scott model with noise (noise strength $\theta = 2 \times 10^{-3}$).

Taking θ equal to zero, Eq. (4.9) reduces to the classical Gray-Scott equation of reaction diffusion system:

$$\begin{pmatrix} \dfrac{\partial u(\mathbf{x},t)}{\partial t} \\ \dfrac{\partial v(\mathbf{x},t)}{\partial t} \end{pmatrix} = \mathbf{M}(u,v) + \mathbf{F}(u,v). \tag{4.13}$$

Then we simulate Eq. (4.13) (in the deterministic case) and Eq. (4.9) (in the stochastic case) using the same initial conditions as used in [124], that the center 0.2×0.2 region takes initial value of $U = 1/2$, $V = 1/4$, on the background of $U = 1$, $V = 0$. These conditions are then perturbed with $\pm 1\%$ random noise to break the square symmetry.

Noise Induced Pattern Formation For the first example, we take the parameters f and k as: $f = 0.053$, $k = 0.06$. The system does not display the phenomenon of pattern formation without noise (as shown in the left panel of Fig. 4.1). When noise is present with a strength of $\theta = 2 \times 10^{-3}$ (under the current choice of the parameters and initial conditions), however, the system forms the dot emerging pattern (shown in the right panel of Fig. (4.1)). The red dots emerges from blue background, similar to the ι patterns observed in the paper [124].

Noise Induced Pattern Change For the second example, we take the parameters f and k as: $f = 0.0483$, $k = 0.06$. Without noise, the system displays the patterns of dot emergence

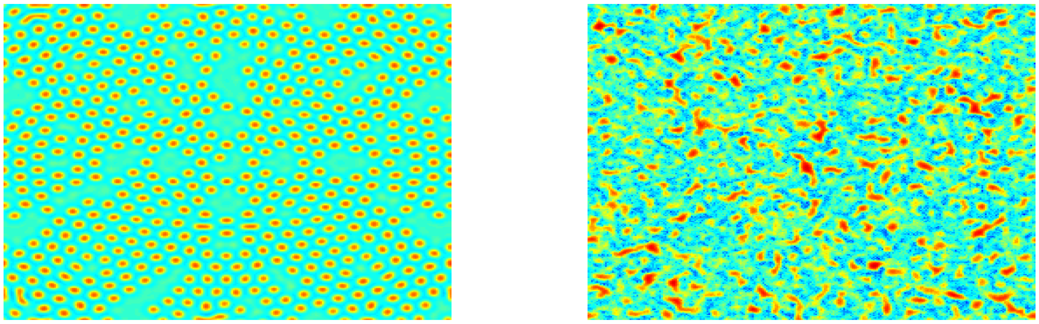


Figure 4.2: Left panel. Long time behavior of the Gray-Scott model without noise ($f = 0.0483$, $k = 0.06$, $M_u = 4 \times 10^{-5}$, $M_v = 2 \times 10^{-5}$); Right panel. Long time behavior of the Gray-Scott model with noise (noise strength $\theta = 3 \times 10^{-3}$).

(as shown in the left panel of Fig. 4.2), like the ones observed in the right panel of Fig. (4.1). When noise is present with a strength of $\theta = 3 \times 10^{-3}$, the patterns become more connected, forming stripes-like structures as the red state becomes more prevalent (shown in the right panel of Fig. 4.2).

4.1.3 Effective Dynamics

In this paper, we are only concerned with fluctuation from chemical reaction. Noise does not influence the diffusion part of the dynamics. Hence we only need to study the effect of noise on the local reaction dynamics $\mathbf{F}(u, v)$ of Eq. (4.9). For the stochastic reaction dynamics:

$$\begin{pmatrix} \mathrm{d}u \\ \mathrm{d}v \end{pmatrix} = \mathbf{F}(u, v)\mathrm{d}t + \theta^{\frac{1}{2}}\sigma(u, v)\mathrm{d}\mathbf{W}(t), \tag{4.14}$$

a Fokker-Planck equation can be written for the evolution of distribution in reactants' density (u, v) at each site:

$$\begin{aligned} \partial_t p(u, v, t) &= \mathcal{L}[p(u, v, t)] \\ &= -\nabla^T \cdot \big(\mathbf{F}(u, v)p(u, v, t)\big) + \theta\nabla^2 : \big(\mathbf{D}(u, v)p(u, v, t)\big). \end{aligned} \tag{4.15}$$

The long-time behavior of the stochastic system as Eq. (4.15) can be characterized by its stationary distribution $p^s(u,v)$. And yet the deterministic vector field $\mathbf{F}(u,v)$ does not correspond to the stationary behavior. A major discrepancy between the original stochastic dynamics Eq. (4.14) and the vector field $\mathbf{F}$ is that the most frequently visited states of the stochastic reaction dynamics are different from the stable attractors of $\mathbf{F}$. This discrepancy can be resolved by focusing on the stochastic dynamics evolving towards stationary behavior.

To focus on the evolution of distribution function $p(u,v,t)$ towards the stationary distribution, we can follow [103] and rewrite Eq. (4.15) as:

$$\partial_t p(u,v,t) = \widetilde{\mathcal{L}}\left[\frac{p(u,v,t)}{p^s(u,v)}\right], \tag{4.16}$$

where operator $\widetilde{\mathcal{L}}$ is:

$$\widetilde{\mathcal{L}}[\varphi(u,v)] = \theta\nabla^T \cdot \Big(\big(\mathbf{D}(u,v) + \mathbf{Q}(u,v)\big)\nabla\varphi(u,v)\cdot p^s(u,v)\Big), \tag{4.17}$$

where $\mathbf{Q}(u,v)$ is a skew-symmetric matrix. Because of the existence of large deviation rate function for the system of Eq. (4.15), we can write $p(u,v,t) = e^{-\phi(u,v,t)/\theta}$; $p^s(u,v) = e^{-\phi^s(u,v)/\theta}$ and have $\phi(u,v,t)$ exists even when $\theta \to 0$. In terms of $\phi(u,v,t)$, we have:

$$\begin{aligned}\frac{\partial\phi(u,v,t)}{\partial t} =& -(\nabla\phi)^T\big(\mathbf{D}(u,v) + \mathbf{Q}(u,v)\big)\nabla\big(\phi(u,v,t) - \phi^s(u,v)\big)\\ &+ \theta\nabla^T\Big(\big(\mathbf{D}(u,v) + \mathbf{Q}(u,v)\big)\nabla\big(\phi(u,v,t) - \phi^s(u,v)\big)\Big)\end{aligned} \tag{4.18}$$

The first term on the right hand side of Eq. (4.18) is of order $\mathcal{O}(1)$ and the second term is of order $\mathcal{O}(\theta)$.

It might be tempting to think that Eq. (4.18) is an expansion result following large deviation theory. However, we should recall that the solution $p(u,v,t)$ to Eq. (4.15) (as long as matrix $\mathbf{Q}(u,v)$) depends on θ and such that $\phi(u,v,t)$ and $\phi^s(u,v)$ actually contain higher order terms of θ. In fact, if we use the WKB ansatz:

$$p(u,v,t) = e^{-\theta^{-1}\phi(u,v,t) + \psi_0(u,v,t) + \theta\,\psi_1(u,v,t) + \cdots} \tag{4.19}$$

to expand Eq. (4.15), we can see that Eq. (4.18) is a *resummation* of all the higher order terms into the first two orders. As a result, Eq. (4.18) is *exact* and when $\phi(u,v,t) \to \phi^s(u,v)$, every term in Eq. (4.18) approaches 0 uniformly. When $\theta \to 0$, the second term on the right hand side of Eq. (4.18) approaches zero and the first term approaches the large deviation rate function for Eq. (4.15).

Most probable trajectory $(u^*(t), v^*(t))$ of Eq. (4.18) for small θ can be found by focusing on the first term on the right hand side of Eq. (4.18) and considering a local minimum of the function $\phi(u^*(t), v^*(t), t)$ located at $(u(t), v(t)) = (u^*(t), v^*(t))$:

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}t}\big(\nabla\phi(u^*(t), v^*(t), t)\big) \\ &= \left[\frac{\partial\nabla\phi(u,v,t)}{\partial t} + \mathbf{H}(\phi(u,v,t))\begin{pmatrix}\dfrac{\mathrm{d}u}{\mathrm{d}t}\\ \dfrac{\mathrm{d}v}{\mathrm{d}t}\end{pmatrix}\right]_{(u(t),v(t))=(u^*(t),v^*(t))}, \end{aligned} \tag{4.20}$$

where $\mathbf{H}(\phi(u,v,t))$ is the Hessian matrix of $\phi(u,v,t)$ over (u,v). Then the most probable trajectories follow the following ODE:

$$\begin{aligned} \begin{pmatrix}\dfrac{\mathrm{d}u^*}{\mathrm{d}t}\\ \dfrac{\mathrm{d}v^*}{\mathrm{d}t}\end{pmatrix} &= -\left[\mathbf{H}^{-1}(\phi(u,v,t))\frac{\partial\nabla\phi(u,v,t)}{\partial t}\right]_{(u(t),v(t))=(u^*(t),v^*(t))} \\ &= -\big(\mathbf{D}(u^*,v^*) + \mathbf{Q}(u^*,v^*)\big)\nabla\phi^s(u^*,v^*). \end{aligned} \tag{4.21}$$

Therefore, the long term effect dynamics follows the vector field:

$$\widetilde{\mathbf{F}}(u,v) = -\left(\mathbf{D}(u,v) + \mathbf{Q}(u,v)\right)\nabla\phi^s(u,v) \tag{4.22}$$

The structure of the effective dynamics $\widetilde{\mathbf{F}}(u,v)$ in Eq. (4.22) ensures that $\phi^s(u,v)$ is the Lyapunov function for $\widetilde{\mathbf{F}}(u,v)$ and that the effective dynamics is always attracted to the mostly distributed states, the same as the long term behavior of the stochastic dynamics. The discrepancy between $\widetilde{\mathbf{F}}(u,v)$ and the deterministic vector field $\mathbf{F}(u,v)$ is:

$$\begin{aligned} \Delta\mathbf{F}_i(u,v) &= \mathbf{F}_i(u,v) - \widetilde{\mathbf{F}}_i(u,v) \\ &= \theta\sum_{j=1,2}\partial_j\left(\mathbf{D}_{ij}(u,v) + \mathbf{Q}_{ij}(u,v)\right). \end{aligned} \tag{4.23}$$

Although The stationary distribution $p^s(u,v)$ is not normalizable and hard to calculate in this case, matrix $\mathbf{Q}(u,v) = \begin{pmatrix} 0 & -q(u,v) \\ q(u,v) & 0 \end{pmatrix}$ can still be calculated by noting that:

$$\nabla\times\left(\mathbf{D}(u,v) + \mathbf{Q}(u,v)\right)^{-1}\mathbf{F}(u,v) = 0. \tag{4.24}$$

An expansion solution for $q(u,v)$ in the interval of $(0,1)\times(0,1)$ can be obtained:

$$
\begin{aligned}
q(u,v) \sim{} & \frac{1}{2}\frac{fk}{f+k} - \frac{1}{2}\frac{fk}{f+k}u + \frac{1}{2}\frac{(-f^3-2f^2k)}{(f+k)^2(3f+2k)}v + \frac{1}{2}\frac{f(f+2k)}{(f+k)^2}uv \\
& + \frac{1}{2}\frac{(-9f^5k+2f^5-39f^4k^2+8f^4k-58f^3k^3+8f^3k^2-36f^2k^4-8fk^5)}{k(4f+3k)\left(3f^2+5fk+2k^2\right)^2}v^2 \\
& + \cdots . \qquad (4.25)
\end{aligned}
$$

Hence, the effective dynamics $\tilde{\mathbf{F}}(u,v)$ can be solved as:

$$
\tilde{\mathbf{F}}(u,v) = \mathbf{F}(u,v) - \Delta\mathbf{F}(u,v). \qquad (4.26)
$$

We find that the effective dynamics is actually closer to the bifurcation point. Simulation of the effective deterministic dynamics with diffusion term $\mathbf{D}(u,v)$ demonstrates that emerging patterns are developed over time.

4.1.4 Noise Induced Dynamic Change

We are now in a position of comparing the dynamic stability of the deterministic reaction diffusion dynamics: $\mathbf{M}(u,v)+\mathbf{F}(u,v)$ and the effective reaction diffusion dynamics: $\mathbf{M}(u,v)+\widetilde{\mathbf{F}}(u,v)$, respectively. We explicitly demonstrate the analysis for the deterministic dynamics $\mathbf{M}(u,v)+\mathbf{F}(u,v)$, and use the same procedure for $\mathbf{M}(u,v)+\widetilde{\mathbf{F}}(u,v)$.

Linear Stability We can first observe that the uniform solution, $\big(u(\mathbf{x}),v(\mathbf{x})\big)=(u_0,v_0)$, where $\mathbf{F}(u_0,v_0)=0$ is a stationary solution for the system:

$$
\begin{pmatrix} \dfrac{\partial u(\mathbf{x},t)}{\partial t} \\ \dfrac{\partial v(\mathbf{x},t)}{\partial t} \end{pmatrix} = \mathbf{M}(u,v)+\mathbf{F}(u,v). \qquad (4.27)
$$

Then we can analyze the linear stability of the system under constant perturbations: $(u(t),v(t)) = (u_0+\epsilon u_1, v_0+\epsilon v_1)$, which boils down to the eigenvalues of the Jacobi matrix:

$$
\begin{aligned}
J\big(\mathbf{F}(u_0,v_0)\big) &= \begin{pmatrix} \dfrac{\partial \mathbf{F}_1(u,v)}{\partial u} & \dfrac{\partial \mathbf{F}_1(u,v)}{\partial v} \\ \dfrac{\partial \mathbf{F}_2(u,v)}{\partial u} & \dfrac{\partial \mathbf{F}_2(u,v)}{\partial v} \end{pmatrix}_{(u,v)=(u_0,v_0)} \\
&= \begin{pmatrix} -v_0^2+f & -2u_0v_0 \\ v_0^2 & 2u_0v_0-(f+k) \end{pmatrix}. \qquad (4.28)
\end{aligned}
$$

When $f > 4(f+k)^2$, a linearly stable uniform solution exists:

$$(u_0, v_0) = \left(\frac{f - \sqrt{f^2 - 4f(f+k)^2}}{2f}, \frac{f + \sqrt{f^2 - 4f(f+k)^2}}{2(f+k)} \right). \tag{4.29}$$

Turing Stability To analyze the stripe patterns generated from the system, we can further analyze Turing stability of the stable uniform solution. A plane wave perturbation: $(u(t), v(t)) = (u_0 + \tilde{u}_{\mathbf{k}}(t), v_0 + \tilde{v}_{\mathbf{k}}(t)) = \left(u_0 + \epsilon\tilde{u}(t)e^{i\mathbf{k}^T\cdot\mathbf{x}}, v_0 + \epsilon\tilde{v}(t)e^{i\mathbf{k}^T\cdot\mathbf{x}}\right)$ can be applied to the system, which boils down to the eigenvalue problem of the following system:

$$\begin{pmatrix} \dfrac{\partial \tilde{u}_{\mathbf{k}}(t)}{\partial t} \\ \dfrac{\partial \tilde{v}_{\mathbf{k}}(t)}{\partial t} \end{pmatrix} = -\mathbf{k}^T\mathbf{k} \begin{pmatrix} M_u\tilde{u}_{\mathbf{k}}(t) \\ M_v\tilde{v}_{\mathbf{k}}(t) \end{pmatrix} + J\big(\mathbf{F}(u_0, v_0)\big) \begin{pmatrix} \tilde{u}_{\mathbf{k}}(t) \\ \tilde{v}_{\mathbf{k}}(t) \end{pmatrix}. \tag{4.30}$$

When one of the eigenvalues of the following matrix, $\widetilde{J}_{\mathbf{k}}$:

$$\widetilde{J}_{\mathbf{k}} = J\big(\mathbf{F}(u_0, v_0)\big) - \mathbf{k}^T\mathbf{k} \begin{pmatrix} M_u & 0 \\ 0 & M_v \end{pmatrix} \tag{4.31}$$

is positive, then the uniform solution is unstable under plane wave perturbation and the stripe patterns would emerge.

Noise Induced Dot Patterns In the first case, we analyze the behavior of the deterministic dynamics and effective dynamics respectively and find that the effective dynamics is closer to the saddle-node bifurcation and that the linearly stable fixed point, "blue state", becomes less stable (real part of the eigenvalues changes from -8.97498×10^{-3} to -5.56142×10^{-3}, losing 40% of the original stability), rendering the dynamics within the regime of causing the localized dot patterns. Although the effective dynamics is still Turing stable (from plan wave perturbations), using techniques discussed in detail in [30] one can find that it is unstable from local perturbations. We also simulate the effective dynamics and compared it against the noisy system. It can be seen from Fig. 4.3 that the effective system's prevailing behaviors is constant dot emergence, the same as the noisy system.

Noise Induced Pattern Change In the second case, we discover that similar to the first case, the effective dynamics is closer to the saddle-node bifurcation and that the the linearly stable fixed point, "blue state", becomes even less stable (real part of the eigenvalues changes from -3.99391×10^{-3} to -5.84034×10^{-4}). More importantly, the "blue state"

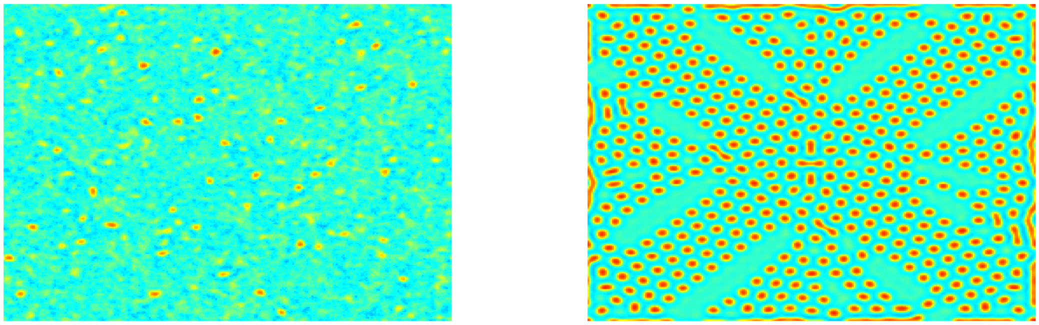


Figure 4.3: Left panel. Dynamics of the Gray-Scott model with noise; Right panel. Long time behavior of the Gray-Scott model under effective dynamics $\mathbf{M} + \widetilde{\mathbf{F}}$ ($f = 0.053$, $k = 0.06$, $M_u = 4 \times 10^{-5}$, $M_v = 2 \times 10^{-5}$).

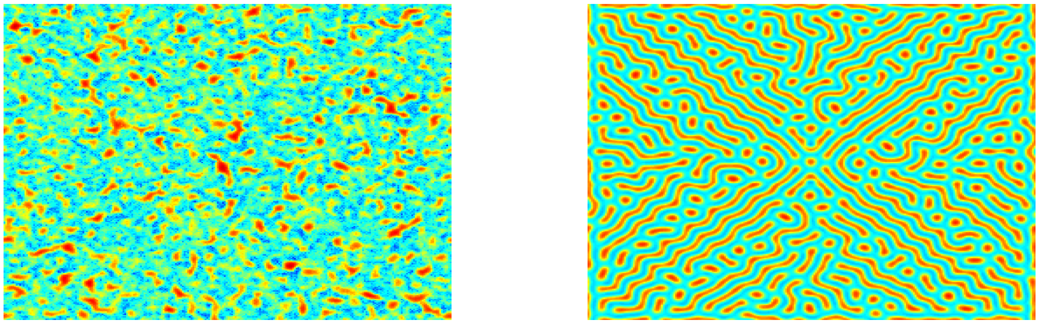


Figure 4.4: Left panel. Dynamics of the Gray-Scott model with noise; Right panel. Long time behavior of the Gray-Scott model under effective dynamics $\mathbf{M} + \widetilde{\mathbf{F}}$ ($f = 0.0483$, $k = 0.06$, $M_u = 4 \times 10^{-5}$, $M_v = 2 \times 10^{-5}$).

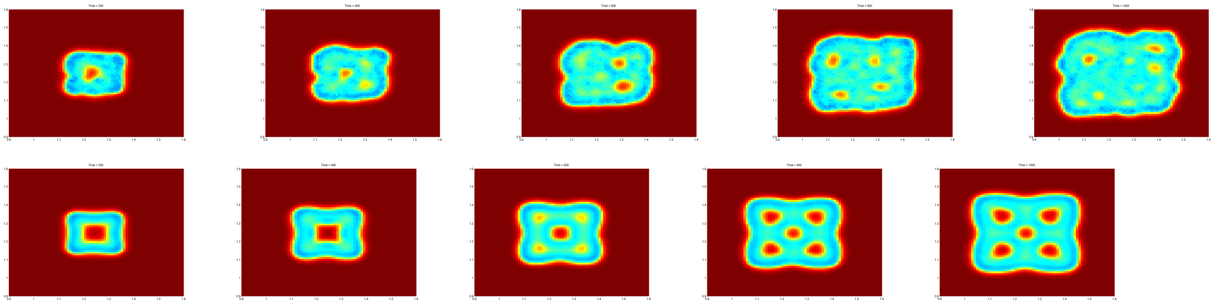


Figure 4.5: Upper panels: Behavior of the stochastic Gray-Scott model over time; Lower panels: Behavior of the Gray-Scott model simulated with the effective deterministic dynamics. Snapshots are taken at the same time for lower and upper panels.

becomes Turing unstable under the effective dynamics, changing the localized dot patterns into continuous stripe patterns. The plane wave perturbations that make the "blue state" unstable have wave numbers ranging from 25 to 45, which is consistent with the width of the stripes simulated in the noisy system. It can be seen from Fig. 4.4 that simulation of the effective dynamics further corroborates with the analytical result.

4.1.5 Discussion

It can be observed from Fig. (4.5) that the behavior of the stochastic Gray-Scott model corresponds to the behavior simulated with the effective deterministic dynamics. Since the behavioral change induced by noise is accounted for by the effective deterministic dynamics, the remaining differences between the stochastic model and the effective deterministic model is caused purely by fluctuation.

The fluctuation here plays two roles through time: one is the constant breaking of spatial symmetry, facilitating the development of complex patterns. The corresponding observation is: patterns develop earlier when in the stochastic model than the effective deterministic model. Another is the destruction of the already developed patterns. The constant perturbation in the concentration field makes large, organized patterns harder to exist (as shown in Fig. (4.3) and Fig. (4.5)).

Chapter 5

A UNIFYING FRAMEWORK FOR DEVISING EFFICIENT AND IRREVERSIBLE MCMC SAMPLERS

Markov chain Monte Carlo (MCMC) methods are the defacto tools for inference in Bayesian models [99, 149]. There are two primary approaches to developing and implementing such MCMC algorithms. One is the traditional Metropolis-Hastings type of approach, where one defines a jump process through an accept-reject procedure. This popular class of methods utilizes global information of the target distribution. Another approach relies on the local (gradient) information of the target distribution and designs a continuous dynamical process with the target distribution as its stationary distribution; samples are proposed according to the integrated trajectories of the continuous dynamics. Important examples of such samplers include Hamiltonian Monte Carlo methods [38, 113] and samplers using Langevin dynamics [150, 184, 182, 121].

For both approaches, a major challenge is devising a sampler with good mixing rates (i.e., the speed at which an initial distribution converges to the target distribution). In the world of jump-process-based MCMC techniques, a focus has been on developing clever proposals [165, 73, 99], but these methods are often strongly coupled to a specific challenge setting, like multimodal targets [165] or heavy tailed distributions [73]. In practice, one often does not know the structure of the target distribution, which might additionally exhibit a combination of these factors. In the world of continuous-dynamic-based samplers, methods using second order information, like Riemannian Hamiltonian Monte Carlo [57], can be helpful. However, it is non-trivial to devise these modifications and prove that the dynamics maintain the right stationary distribution.

In this paper, we propose a unifying framework for these two approaches that enables a more user-friendly method for devising efficient and general-purpose MCMC procedures. We start by examining continuous dynamic samplers. We present a stochastic differential equation (SDE) based framework in which to define all such valid samplers based on specifying two matrices: a positive semidefinite diffusion matrix and a skew-symmetric curl matrix. These matrices define symmetric (reversible) and anti-symmetric (irreversible) operators, respectively. Based on this framework, we prove that for any choice of these matrices, the sampler will have the target distribution as the stationary distribution. We likewise prove that any continuous dynamic sampler with the correct stationary distribution has a representation in this framework. As such, we call this framework *complete*. We cast a number of past methods in our proposed representation, and also show how it can be used to devise new samplers with improved mixing rates. An initial version of this work appeared in [102].

In [102], jump processes were specifically excluded from the analysis. However, jump processes represent a potentially attractive approach to MCMC since, in theory, samples generated from jump processes can decorrelate rapidly. In practice, it is challenging to define transition kernels that enable this efficient exploration while maintaining a reasonably high acceptance rate. The challenge partially stems from the fact that attention has typically been restricted to *reversible* jump processes. Such samplers are straightforward to derive and implement, but the reversibility restriction hinders the mixing rates and efficiency of the proposed algorithms [112, 36, 29, 27]. Unfortunately, devising irreversible samplers is a non-trivial task, and often results in computationally complex algorithms.

Leveraging our insights from the operator decomposition for continuous dynamic processes, we show that a similar framework can enable the development of efficient *irreversible* jump process samplers based on the specification of two kernel functions. We decompose the jump operator into symmetric (reversible) and anti-symmetric (irreversible) operators, paralleling the continuous-dynamic sampler framework. Using this decomposition and parameterization, we arise at a straightforward set of constraints on the transition kernels that ensures that the target distribution is the stationary distribution. The resulting sampler implementation has the ease and efficiency of Metropolis-Hastings; in fact, the implementation directly parallels that of standard Metropolis-Hastings. In terms of runtime, our proposed method significantly outperform previous approaches [107, 63], in addition to providing fast mixing rates in a range of scenarios, from heavy-tailed to multimodal targets. We demonstrate these performance gains against existing approaches in a variety of sampling tasks.

There are many ways we can think of combining our continuous dynamic and jump process frameworks. One is to use the continuous dynamic sampler and jump process sampler iteratively, i.e., use one (possibly for multiple iterations) and then the other, just as in Hamiltonian Monte Carlo. Another approach is to use the continuous dynamic sampler for some variables (e.g., real-valued variables) and the jump process sampler for others (e.g., discrete-valued variables). It is straightforward to combine our approaches in these manners since each process maintains the correct stationary distribution, so these types of compositions will likewise result in the correct stationary distribution under some mild conditions at stationary. The strategy of alternating between continuous dynamics and jump processes is similar to what was recently proposed in the *bouncy particle* [21] and *Zig-Zag* [17, 16] samplers. These samplers iterate between deterministic continuous dynamics and Poisson jump processes. However, the algorithms associated with these methods are quite involved and deviate significantly from classical MCMC tools.

Alternatively, as we show, we are able to use a discretization of the continuous dynamics as a proposal distribution in our jump process accept-reject scheme, even when the continuous dynamics are not reversible. Here, the transition kernel is defined according to the SDE representation of the continuous dynamics. Importantly, the simplicity of the Metropolis-Hastings algorithm is inherited. We can view the benefits of this approach from two angles: (i) the SDE can provide an efficient proposal distribution for our irreversible Markov jump process or (ii) the accept-reject scheme allows us to correct for the bias introduced by sampling the continuous dynamics via a discretized SDE. This accept-reject scheme is a direct generalization of the Metropolis-adjusted Langevin algorithm (MALA) [150] to irreversible SDEs. This opens up the possibility to combine, for example, the benefits of Langevin diffusion with Hamiltonian dynamics. As we see, combining our frameworks for continuous dynamic processes and Markov jump processes yields a unified and complete framework for devising efficient and general purpose MCMC samplers with the correct stationary distribution.

We conclude with a discussion on how to scale our continuous and jump frameworks to perform Bayesian inference in large datasets.

5.1 Backgrounds and Standard Approach

We start with the standard MCMC goal of drawing samples from a target distribution $\pi(\mathbf{z})$. The idea behind MCMC is to translate the task of sampling from the posterior distribution to simulating from a Markov process. One can then discuss the evolution of the distribution on $\mathbf{z}$ at time t, $p(\mathbf{z}; t)$, under this stochastic process and analyze its *stationary distribution*, $p^s(\mathbf{z})$. If the stochastic process is ergodic and its stationary distribution is equal to the target distribution $\pi(\mathbf{z})$, then simulating the stationary stochastic dynamics equates with providing samples from the posterior distribution.

In this section, we review some of the fundamentals of the stochastic processes associated with general Markov processes, and how we can use these processes to construct samplers.

5.1.1 Backgrounds on Constructing Samplers from Continuous and Jump Markov Processes

In Section 2.1.2, we described that the realization of the continuous Markov process can be generated from the stochastic differential equation (SDE):

$$\mathrm{d}\mathbf{z} = \mathbf{f}(\mathbf{z})\mathrm{d}t + \sqrt{2\mathbf{D}(\mathbf{z})}\mathrm{d}\mathbf{W}(t). \tag{5.1}$$

With an abuse of notation, in this and the following chapter, we follow conventions in the classical MCMC literature and use the same boldfaced lowercase letter (e.g., $\mathbf{z}$) to denote both a random vector and its possible value in $\mathbb{R}^d$. In practice, an ϵ-discretization is used to simulate from (5.1):

$$\mathbf{z}_{t+1} \leftarrow \mathbf{z}_t + \epsilon_t \mathbf{f}(\mathbf{z}_t) + \eta_t, \quad \eta_t \sim \mathcal{N}(0, 2\epsilon_t \mathbf{D}(\mathbf{z}_t)) \tag{5.2}$$

Although (5.2) is in the form of the Euler–Maruyama method, higher order numerical schemes can be used for better accuracy [26, 20, 94]. The challenge here is to select $\mathbf{f}$ and $\mathbf{D}$ such that the simulation from (5.1) leads to the right stationary distribution. Note that relying on a sample path from the discretized system of (5.2) typically leads to the introduction of bias due to discretization error. In these cases, the samples only provide unbiased estimates in the limit as $\epsilon_t \to 0$ unless further corrections are introduced. In Section 5.2, using the decomposition idea from Sec. 2.3, we propose a reparameterization of (5.1) from which it is trivial to ensure the SDE has the desired stationary distribution. Then, in Section 5.4, we return to the idea of correcting for any potential discretization error if no bias can be tolerated.

Turning to the Markov jump process, a realization of it (as discussed in Sec. 2.1.2) can be implemented by using the following transition probability:

$$p(\mathbf{z}|\mathbf{y};\Delta t) = \Delta t W(\mathbf{z}|\mathbf{y}) + \left[1 - \Delta t \int_{\mathbb{R}^d} W(\mathbf{x}|\mathbf{y})\mathrm{d}\mathbf{x}\right] \delta(\mathbf{z} - \mathbf{y}). \tag{5.3}$$

Equation (5.3) corresponds to a sampling process as follows. We take Δt to be the stepsize. Then, with probability $\Delta t W(\mathbf{z}|\mathbf{y})$, we transit from state $\mathbf{y}$ to state $\mathbf{z}$. With probability $1 - \Delta t \int_{\mathbb{R}^d} W(\mathbf{x}|\mathbf{y})\mathrm{d}\mathbf{x}$, we stay in state $\mathbf{y}$.

Analogously to the challenge of selecting $\mathbf{f}$ and $\mathbf{D}$, the challenge here is to select the kernel $W(\mathbf{z}|\mathbf{x})$ that leads to the stationary distribution being equal to the target distribution, $\pi(\mathbf{z})$. This requires that the positive transition kernel W satisfies

$$\int_{\mathbb{R}^d} \mathrm{d}\mathbf{x} \left[W(\mathbf{z}|\mathbf{x})\pi(\mathbf{x}) - W(\mathbf{x}|\mathbf{z})\pi(\mathbf{z})\right] = 0.$$

Additionally, even if such a W can be defined, it might not define a distribution from which we can straightforwardly sample nor compute the necessary integral.

5.1.2 Standard Metropolis Hastings (MH) Algorithm

Instead, in practice one typically resorts to implementing jump process samplers through the Metropolis-Hastings (MH) accept-reject scheme [107, 63]. In MH (Algorithm 1), one samples from a specified *proposal distribution* $q(\mathbf{z}|\mathbf{y})$ and accepts the proposed value $\mathbf{z}$ with probability

$$\alpha(\mathbf{y},\mathbf{z}) = \min\left(1, \frac{\pi(\mathbf{z})q(\mathbf{y}|\mathbf{z})}{\pi(\mathbf{y})q(\mathbf{z}|\mathbf{y})}\right). \tag{5.4}$$

In the form of (2.10), we have [31]:

$$p(\mathbf{z}|\mathbf{y};\Delta t) = q(\mathbf{z}|\mathbf{y})\alpha(\mathbf{y},\mathbf{z}) + \left[1 - \int_{\mathbb{R}^d} q(\mathbf{x}|\mathbf{y})\alpha(\mathbf{y},\mathbf{x})\mathrm{d}\mathbf{x}\right]\delta(\mathbf{z}-\mathbf{y}). \tag{5.5}$$

When in state $\mathbf{y}$ at time t, we propose to jump to state $\mathbf{z}$ at $t+\Delta t$ with conditional probability $q(\mathbf{z}|\mathbf{y})$, realized via a random number generator that has a distribution according to $q(\mathbf{z}|\mathbf{y})$; we accept this proposal with probability $\alpha(\mathbf{y},\mathbf{z})$ to ensure that the target distribution will be preserved under this procedure. Hence, the total probability of transiting from state $\mathbf{y}$ to $\mathbf{z}$ is $q(\mathbf{z}|\mathbf{y})\alpha(\mathbf{y},\mathbf{z})$. Otherwise, we stay in state $\mathbf{y}$.

Comparing (5.5) to (5.3), we see that MH restricts our attention to $W(\mathbf{z}|\mathbf{y})$ satisfying $\Delta t W(\mathbf{z}|\mathbf{y}) = q(\mathbf{z}|\mathbf{y})\alpha(\mathbf{y},\mathbf{z})$. We further see that the acceptance rate $\alpha(\mathbf{y},\mathbf{z})$ is specifically designed so that $W(\mathbf{z}|\mathbf{x})\pi(\mathbf{x}) = W(\mathbf{x}|\mathbf{z})\pi(\mathbf{z})$, a condition much stronger than $\int_{\mathbb{R}^d}\mathrm{d}\mathbf{x}\,[W(\mathbf{z}|\mathbf{x})\pi(\mathbf{x}) - W(\mathbf{x}|\mathbf{z})\pi(\mathbf{z})] = 0$, in order to ensure that $\pi(\mathbf{z})$ is the stationary distribution. However, this form restricts our attention solely to reversible processes. Instead, just as in the continuous Markov process case, in Section 5.3 we consider a reparameterization in terms of two kernel functions with straightforward-to-satisfy constraints. In that form, not only do we ensure the right stationary distribution, but we are able to devise a sampling algorithm for *irreversible* processes that has the same simplicity of implementation as MH.

As we will show in Sections 5.2 and 5.3, via a reparameterization of the continuous and jump Markov processes (Eq. (5.1) and (5.3)), we transform the problem of devising continuous and jump samplers with the right stationary distribution to one of simply specifying two matrices and two modestly constrained kernel functions. We can compose these two processes in various ways and still ensure that the overall sampler has the correct stationary distribution as is discussed in Section 5.4.

Algorithm 1: Metropolis-Hastings Algorithm

```
for $t = 0, 1, 2 \cdots N_{iter}$ do
    sample $u \sim \mathcal{U}_{[0,1]}$
    propose $\mathbf{z}(*) \sim q(\mathbf{z}(*)|\mathbf{z}(t))$
    $\alpha\left(\mathbf{z}^{(t)}, \mathbf{z}(*)\right) = \min\left\{1, \frac{\pi\left(\mathbf{z}(*)\right) q(\mathbf{z}(t)|\mathbf{z}(*))}{\pi\left(\mathbf{z}(t)\right) q(\mathbf{z}(*)|\mathbf{z}(t))}\right\}$
    if $u < \alpha\left(\mathbf{z}(t), \mathbf{z}(*)\right)$, $\mathbf{z}(t+1) = \mathbf{z}(*)$
    else $\mathbf{z}(t+1) = \mathbf{z}(t)$
end
```

5.2 Continuous Markov Process Based Samplers

As proposed in Sec. 2.3, Eq. (5.1) can be reparameterized in the following form:

$$\mathrm{d}\mathbf{z} = \Big[-\big(\mathbf{D}(\mathbf{z}) + \mathbf{Q}(\mathbf{z})\big)\nabla H(\mathbf{z}) + \Gamma(\mathbf{z})\Big]\mathrm{d}t + \sqrt{2\mathbf{D}(\mathbf{z})}\mathrm{d}\mathbf{W}(t), \tag{5.6}$$

where $H(\mathbf{z}) = -\log(\pi(\mathbf{z}))$ and $\Gamma_i(\mathbf{z}) = \sum_{j=1}^{d} \frac{\partial}{\partial \mathbf{z}_j}\big(\mathbf{D}_{ij}(\mathbf{z}) + \mathbf{Q}_{ij}(\mathbf{z})\big)$. For any positive semi-definite $\mathbf{D}$ and skew-symmetric $\mathbf{Q}$ matrices, invariant distribution of Eq. (5.6) is the desired target distribution $\pi(\mathbf{z})$. The problem of finding the correct Markov dynamics has just been translated to choosing from all possible matrices $\mathbf{D}$ and $\mathbf{Q}$. Furthermore, because any continuous Markov process has a representation in the form of Eq. (5.6), the recipe parameterized by $\mathbf{D}$ and $\mathbf{Q}$ is *complete*.

Following (2.9), we can simulate from the SDE in Eq. (5.6) using the following ϵ-discretization:

$$\mathbf{z}_{t+1} \leftarrow \mathbf{z}_t - \epsilon_t \left[\big(\mathbf{D}(\mathbf{z}_t) + \mathbf{Q}(\mathbf{z}_t)\big)\nabla H(\mathbf{z}_t) + \Gamma(\mathbf{z}_t)\right] + \eta_t, \quad \eta_t \sim \mathcal{N}(0, 2\epsilon_t \mathbf{D}(\mathbf{z}_t)). \tag{5.7}$$

Again, higher-order numerical schemes can be used in place of the first-order integrator above [26, 20, 94]. The resulting algorithm is outlined in Algorithm 2. (Recall that bias is introduced via the discretization, i.e., setting ϵ_t finite. We will return to this in Section 5.4.)

5.2.1 Previous MCMC Algorithms as Special Cases

We explicitly state how some previous continuous dynamics used in the MCMC methods fit within the proposed framework based on specific choices of $\mathbf{D}(\mathbf{z})$, $\mathbf{Q}(\mathbf{z})$ and $H(\mathbf{z})$. We show how our framework can be used to "reinvent" the samplers by guiding their construction and avoiding potential mistakes or inefficiencies caused by naïve implementations.

Hamiltonian Monte Carlo (HMC) The key ingredient in HMC [38, 113] is Hamiltonian dynamics, which models the physical motion of an object with position θ, momentum r, and mass $\mathbf{M}$ on an frictionless surface following the potential well $U(\theta) = -\log \pi(\theta)$ as follows:

$$\begin{cases} \mathrm{d}\theta = \mathbf{M}^{-1} r \mathrm{d}t \\ \mathrm{d}r = -\nabla U(\theta_t)\mathrm{d}t. \end{cases} \tag{5.8}$$

Equation (6.8) is a special case of the proposed framework with $\mathbf{z} = (\theta, r)$, $H(\theta, r) = U(\theta) + \frac{1}{2} r^T \mathbf{M}^{-1} r$, $\mathbf{Q}(\theta, r) = \begin{pmatrix} 0 & -\mathbf{I} \\ \mathbf{I} & 0 \end{pmatrix}$ and $\mathbf{D}(\theta, r) = \mathbf{0}$. When $\mathbf{M}$ is taken to be adaptive (in particular, to approximate the Hessian of the target distribution), the method is called Riemannian HMC [57].

Langevin Dynamics The Langevin dynamics sampler [150, 182] proposes to use the following first order (no momentum) Langevin dynamics to generate samples

$$\mathrm{d}\theta = -\mathbf{D}\nabla U(\theta_t)\mathrm{d}t + \sqrt{2\mathbf{D}}\, \mathrm{d}\mathbf{W}(t), \tag{5.9}$$

This algorithm corresponds to taking $\mathbf{z} = \theta$ with $H(\theta) = U(\theta) = -\log \pi(\theta)$, $\mathbf{D}(\theta) = \mathbf{D}$, and $\mathbf{Q}(\theta) = \mathbf{0}$.

Riemannian Langevin Dynamics The Langevin dynamics sampler can be generalized to use an adaptive diffusion matrix $\mathbf{D}(\theta)$. Specifically, it is interesting to take $\mathbf{D}(\theta) = \mathbf{G}^{-1}(\theta)$, where $\mathbf{G}(\theta)$ is the Fisher information metric [184, 121]. The sampler iterates

$$\mathrm{d}\theta = -[\mathbf{G}(\theta)^{-1}\nabla U(\theta) + \Gamma(\theta)]\mathrm{d}t + \sqrt{2\mathbf{G}(\theta)^{-1}}\, \mathrm{d}\mathbf{W}(t). \tag{5.10}$$

We can cast this Riemannian Langevin dynamics sampler [121] into our framework taking $\mathbf{D}(\theta) = \mathbf{G}(\theta)^{-1}$, and $\mathbf{Q}(\theta) = \mathbf{0}$. From our framework, we know that here $\Gamma_i(\theta) = \sum_j \frac{\partial \mathbf{D}_{ij}(\theta)}{\partial \theta_j}$. Interestingly, in earlier literature [57], $\Gamma_i(\theta)$ was taken to be

$$2\, |\mathbf{G}(\theta)|^{-1/2} \sum_j \frac{\partial}{\partial \theta_j} \left(\mathbf{G}_{ij}^{-1}(\theta) |\mathbf{G}(\theta)|^{1/2} \right).$$

Algorithm 2: Continuous Markov Process Sampling Algorithm

```
initialize $\mathbf{z}_0$
for $t = 0, 1, 2 \cdots N_{iter}$ do
    for $i = 1 \cdots n$ do
        $\Gamma_i(\mathbf{z}) = \sum_j \frac{\partial}{\partial \mathbf{z}_j} (\mathbf{D}_{ij}(\mathbf{z}) + \mathbf{Q}_{ij}(\mathbf{z}))$
    end
    sample $\eta_t \sim \mathcal{N}(0, 2\epsilon_t \mathbf{D}(\mathbf{z}_t))$
    $\mathbf{z}_{t+1} \leftarrow \mathbf{z}_t - \epsilon_t \left[\big(\mathbf{D}(\mathbf{z}_t) + \mathbf{Q}(\mathbf{z}_t)\big)\nabla H(\mathbf{z}_t) + \Gamma(\mathbf{z}_t)\right] + \eta_t$
end
```

More recently, it was found that this correction term corresponds to the distribution function with respect to a non-Lebesgue measure [150]. For the Lebesgue measure, the revised $\Gamma_i(\theta)$ was as determined by our framework [150]. This is an example of how our theory provides guidance in devising correct samplers.

Summary of past samplers In our framework, the Langevin dynamic based samplers take $\mathbf{Q}(\mathbf{z}) = 0$ and instead stress the design of the diffusion matrix $\mathbf{D}(\mathbf{z})$. The standard Langevin dynamic sampler uses a constant $\mathbf{D}(\mathbf{z})$, whereas the Riemannian variant uses an adaptive, $\mathbf{z}$-dependent diffusion matrix to better account for the geometry of the space being explored. On the other hand, HMC takes $\mathbf{D}(\mathbf{z}) = 0$ and focuses on the curl matrix $\mathbf{Q}(\mathbf{z})$. As we see, our method is a generalization of the dynamics underlying Hamiltonian Monte Carlo (and Riemannian Hamiltonian Monte Carlo) methods, extending the symplectic structure to be non-constant. That is, considering general skew-symmetric matrices instead of just $\mathbf{Q}(\mathbf{z}) = \begin{pmatrix} 0 & -\mathbf{I} \\ \mathbf{I} & 0 \end{pmatrix}$ as in [170, 42]. The generalized Hamiltonian dynamics can explore the state space rapidly, and are guaranteed to preserve the stationary distribution once it is achieved.

Examination of the past methods provides us with the insight that $\mathbf{D}(\mathbf{z})$ can enable diffusive exploration across local modes. And just as in HMC, $\mathbf{Q}(\mathbf{z})$ drives the sampler to walk along contours of equal probability allowing it to rapidly traverse regions of lower probability, especially when state adaptation is incorporated. Importantly, through our $(\mathbf{D}(\mathbf{z}), \mathbf{Q}(\mathbf{z}))$ parameterization, we can readily examine which parts of the product space

$\mathbf{D}(\mathbf{z}) \times \mathbf{Q}(\mathbf{z})$—representing the space of all possible samplers—have been covered. We see that a majority of possible samplers have not yet been considered. For ways in which to use our framework to construct new samplers, see [102].

5.3 Markov Jump Process Based Samplers

Similar to the continuous-dynamic-based MCMC methods, the form of (5.3) specified by the generic kernel $W(\mathbf{x}|\mathbf{z})$ poses challenges to determining the choices of $W(\mathbf{x}|\mathbf{z})$ that lead to a jump process with the correct stationary distribution. Even if one can construct such a W, it can be challenging to use W to sample a realization of the jump process; instead, often one restricts attention to reversible processes and uses MH (see Section 5.1.1). We instead turn to Sec. 2.4 for the following equivalent but alternative representation defined in terms of the two kernels S and A:

$$\frac{\partial p(\mathbf{z}|\mathbf{y};t)}{\partial t} = \int_{\mathbb{R}^d} \mathrm{d}\mathbf{x} \left[\big(S(\mathbf{x},\mathbf{z}) + A(\mathbf{x},\mathbf{z})\big)\varphi(\mathbf{x}) - S(\mathbf{x},\mathbf{z})\varphi(\mathbf{z}) \right], \tag{5.11}$$

where S is a symmetric kernel and A is an anti-symmetric kernel. Based on the form of (5.11), we notice that the requirement that $\pi(\mathbf{z})$ is a stationary distribution of the jump process is translated into simpler constraints: $\int_{\mathbb{R}^d} S(\mathbf{x},\mathbf{z})\pi^{-1}(\mathbf{x})\mathrm{d}\mathbf{x}$ and $\int_{\mathbb{R}^d} A(\mathbf{x},\mathbf{z})\pi^{-1}(\mathbf{x})\mathrm{d}\mathbf{x}$ exists, with $S(\mathbf{x},\mathbf{z}) + A(\mathbf{x},\mathbf{z}) > 0$, and $\int_{\mathbb{R}^d} A(\mathbf{x},\mathbf{z})\mathrm{d}\mathbf{x} = 0$. This enables more ready analysis of the properties of the process, and the development of efficient irreversible jump process samplers.

Following (5.3), the transition probability implied by Eq. (5.11) assuming a Δt-discretization is given by:

$$p(\mathbf{z}|\mathbf{y};\Delta t) = \frac{\Delta t}{\pi(\mathbf{y})}\big(S(\mathbf{y},\mathbf{z}) + A(\mathbf{y},\mathbf{z})\big) + \left[1 - \frac{\Delta t}{\pi(\mathbf{y})}\int_{\mathbb{R}^d} S(\mathbf{y},\mathbf{x})\mathrm{d}\mathbf{x}\right]\delta(\mathbf{z}-\mathbf{y}). \tag{5.12}$$

5.3.1 Previous Samplers as Special Cases

As with past continuous-dynamic-based samplers, we now cast a set of past jump-process-based samplers into our framework.

Direct resampling Methods that sample directly from $\pi(\mathbf{z})$ take $S(\mathbf{y},\mathbf{z}) = \dfrac{1}{\Delta t}\pi(\mathbf{y})\pi(\mathbf{z})$ and $A(\mathbf{y},\mathbf{z}) = 0$. We can verify this by substituting into (2.50).

Metropolis-Hastings The very popular MH algorithm falls into our framework taking $A(\mathbf{y},\mathbf{z}) = 0$ and

$$S(\mathbf{y},\mathbf{z}) = \frac{1}{\Delta t}\min\big(\pi(\mathbf{y})q(\mathbf{z}|\mathbf{y}), \pi(\mathbf{z})q(\mathbf{y}|\mathbf{z})\big). \tag{5.13}$$

To see this, we refer to Section 5.1.1. The specified form for S and A above arises from comparing the transition probability of (5.5) with that of (5.12).

Summary of past samplers In the previously mentioned algorithms, and a majority of those used in practice, only reversible Markov jump processes ($A(\mathbf{z},\mathbf{y}) = 0$) are considered. In Section 5.3.2, we explore the case where the process is irreversible, i.e., $A(\mathbf{z},\mathbf{y}) \neq 0$.

5.3.2 Irreversible Jump Sampler

Analogous to the discussion of Section 5.1.1, there are two issues of designing samplers with Markov jump processes. One is the construction of transition kernels, a task that has been alleviated in part by the new formulation of (5.12) in terms of $S(\mathbf{y},\mathbf{z})$ and $A(\mathbf{y},\mathbf{z})$ with simple constraints, though we still have to construct such kernels. Another is simulating the Markov process of (5.12). In all but the simplest cases, we might not be able to sample from the transition probability $\Delta t \cdot (S(\mathbf{y},\mathbf{z}) + A(\mathbf{y},\mathbf{z}))/\pi(\mathbf{y})$. These two issues are often intertwined posing challenges to the design of samplers. As mentioned in Section 5.1.1, the MH algorithm is often resorted to due to its ease of implementation. It separates the process of proposing a sample into two simple steps: (1) proposing a candidate according to a known conditional probability distribution $q(\mathbf{z}|\mathbf{y})$ and (2) accepting or rejecting the candidate according to a certain probability. An important drawback of the vanilla MH sampler, however, is that the reversibility of the jump process being designed can greatly restrict possible ways to increase the mixing of the Markov chain.

There have been previous efforts to break the restriction of reversibility in different cases. For example, the *non-reversible MH* algorithm adds a vorticity function to the MH procedure [15] while the *lifting method* makes two replica of the original state space with a skew detailed balance condition to facilitate irreversibility [172, 176]. The authors have shown examples of sampling special distributions, but it is unclear how to generalize these past methods to handle a broad set of target distributions. See Section 5.5 for a detailed discussion of these and other methods. Here, we show how we can devise a practical and efficient irreversible jump process algorithm analogous to MH that can be applied to general targets; this procedure implicitly defines valid kernels $S(\mathbf{y},\mathbf{z})$ and $A(\mathbf{y},\mathbf{z})$. In particular, just as MH corresponds to restricting the class of kernels $W(\mathbf{z}|\mathbf{y})$, our algorithm also focuses in on particular instances of $A(\mathbf{y},\mathbf{z})$, but importantly allows $A(\mathbf{y},\mathbf{z}) \neq 0$ (i.e., irreversible processes). The value of this in practice is demonstrated in the experiments of Section 5.3.3.

A naïve approach A straightforward approach to revise the MH algorithm to make anti-symmetric kernel $A(\mathbf{y},\mathbf{z})$ nonzero, resulting in an irreversible sampler, is to utilize different proposal distributions $f(\mathbf{z}|\mathbf{y})$ and $g(\mathbf{z}|\mathbf{y})$, instead of a single $q(\mathbf{z}|\mathbf{y})$. That is, the transition kernel of the MH algorithm in (5.13) is changed to

$$F(\mathbf{y},\mathbf{z}) = S(\mathbf{y},\mathbf{z}) + A(\mathbf{y},\mathbf{z}) = \frac{1}{\Delta t} \min\big(\pi(\mathbf{y})f(\mathbf{z}|\mathbf{y}), \pi(\mathbf{z})g(\mathbf{y}|\mathbf{z})\big). \tag{5.14}$$

Here we are considering jump processes with $A(\mathbf{y},\mathbf{z}) = \frac{1}{2}\big(F(\mathbf{y},\mathbf{z}) - F(\mathbf{z},\mathbf{y})\big) \neq 0$, in contrast to what we saw for MH. By adjusting f and g, faster mixing rates can possibly be attained while maintaining a simple sampling procedure akin to that of MH (see Algorithm 1, but with f in place of q in the numerator and g in place of q in the denominator in the α calculation). The more f and g differ, the more irreversibility effect is incorporated in the design of the sampler. Functions f and g can even be selected to have non-symmetric support in the state space (as is chosen in our experiments), so that new proposals are guided in certain directions until being rejected, encouraging the algorithm to explore farther states. The primary issue with this construction is that $\int_{\mathbb{R}^d} A(\mathbf{y},\mathbf{z})\mathrm{d}\mathbf{y} \neq 0$ in general, rendering the stationary distribution *not* the $\pi(\mathbf{z})$ that we desire. The question is how to design the anti-symmetric kernel $A(\mathbf{y},\mathbf{z})$, such that $\int_{\mathbb{R}^d} A(\mathbf{y},\mathbf{z})\mathrm{d}\mathbf{y} = 0$.

Lifting for sampling when $d = 1$ A simple modified approach is to follow an adjoint Markov process after being rejected by the original one. This is inspired by the *lifting* idea in discrete spaces [172, 176]. Importantly, this approach has $\pi(\mathbf{z})$ as the stationary distribution.

Algorithmically, this process introduces a one-dimensional, uniformly distributed discrete auxiliary variable $\mathbf{y}^p \in \{-1, 1\}$. We then define

$$\begin{aligned} \widetilde{f}(\mathbf{z},\mathbf{z}^p|\mathbf{y},\mathbf{y}^p) &= \big(\mathcal{H}(\mathbf{y}^p)f(\mathbf{z}|\mathbf{y}) + \mathcal{H}(-\mathbf{y}^p)g(\mathbf{z}|\mathbf{y})\big) \\ \widetilde{g}(\mathbf{z},\mathbf{z}^p|\mathbf{y},\mathbf{y}^p) &= \big(\mathcal{H}(-\mathbf{y}^p)f(\mathbf{z}|\mathbf{y}) + \mathcal{H}(\mathbf{y}^p)g(\mathbf{z}|\mathbf{y})\big), \end{aligned} \tag{5.15}$$

where $f(\mathbf{z}|\mathbf{y})$ and $g(\mathbf{z}|\mathbf{y})$ are different conditional probability distributions, and $\mathcal{H}$ is the

Algorithm 3: One-Dimensional Irreversible Jump Sampler

```
randomly pick $z^p$ from $\{1, -1\}$ with equal probability
for $t = 0, 1, 2 \cdots N_{iter}$ do
    sample $u \sim \mathcal{U}_{[0,1]}$
    if $z^p > 0$ then
        sample $\mathbf{z}(*) \sim f\left(\mathbf{z}(*)|\mathbf{z}(t)\right)$
        $\alpha\left(\mathbf{z}(t), \mathbf{z}(*)\right) = \min\left\{1, \frac{\pi\left(\mathbf{z}(*)\right) g\left(\mathbf{z}(t)|\mathbf{z}(*)\right)}{\pi\left(\mathbf{z}(t)\right) f\left(\mathbf{z}(*)|\mathbf{z}(t)\right)}\right\}$
    end
    else
        sample $\mathbf{z}(*) \sim g\left(\mathbf{z}(*)|\mathbf{z}(t)\right)$
        $\alpha\left(\mathbf{z}(t), \mathbf{z}(*)\right) = \min\left\{1, \frac{\pi\left(\mathbf{z}(*)\right) f\left(\mathbf{z}(t)|\mathbf{z}(*)\right)}{\pi\left(\mathbf{z}(t)\right) g\left(\mathbf{z}(*)|\mathbf{z}(t)\right)}\right\}$
    end
    if $u < \alpha\left(\mathbf{z}(t), \mathbf{z}(*)\right)$, $\mathbf{z}(t+1) = \mathbf{z}(*)$; $z^p(t+1) = z^p(t)$
    else $\mathbf{z}(t+1) = \mathbf{z}(t)$; $z^p(t+1) = -z^p(t)$
end
```

Heaviside function:

$$\mathcal{H}(\mathbf{y}^p) = \begin{cases} 1 & \mathbf{y}^p \geq 0 \\ 0 & \mathbf{y}^p < 0. \end{cases}$$

We modify the MH algorithm as described in Algorithm 3, where we update state $\mathbf{y}$ and the auxiliary variable $\mathbf{y}^p$ according to the following transition probability (as in our recipe of (2.50)):

$$\begin{aligned} p(\mathbf{z}, \mathbf{z}^p|\mathbf{y}, \mathbf{y}^p; \Delta t) = & \frac{\Delta t}{\pi(\mathbf{y})\pi(\mathbf{y}^p)}\delta(\mathbf{z}^p - \mathbf{y}^p) \cdot \mathfrak{F}(\mathbf{y}, \mathbf{y}^p, \mathbf{z}, \mathbf{z}^p) \\ & + \delta(\mathbf{z}^p + \mathbf{y}^p)\delta(\mathbf{z} - \mathbf{y})\left(1 - \frac{\Delta t}{\pi(\mathbf{y})\pi(\mathbf{y}^p)}\int_{\mathbb{R}^d} \mathfrak{F}(\mathbf{y}, \mathbf{y}^p, \mathbf{x}, -\mathbf{z}^p)\mathrm{d}\mathbf{x}\right), \end{aligned} \tag{5.16}$$

in which $\mathfrak{F}(\mathbf{y}, \mathbf{y}^p, \mathbf{z}, \mathbf{z}^p)$ is defined using $\widetilde{f}$ and $\widetilde{g}$:

$$\mathfrak{F}(\mathbf{y}, \mathbf{y}^p, \mathbf{z}, \mathbf{z}^p) = \min\left(\pi(\mathbf{y})\pi(\mathbf{y}^p)\widetilde{f}(\mathbf{z}, \mathbf{z}^p|\mathbf{y}, \mathbf{y}^p), \pi(\mathbf{z})\pi(\mathbf{z}^p)\widetilde{g}(\mathbf{y}, \mathbf{y}^p|\mathbf{z}, \mathbf{z}^p)\right).$$

This update rule can be understood as follows. With probability $\mathfrak{F}(\mathbf{y},\mathbf{y}^p,\mathbf{z},\mathbf{z}^p)/(\pi(\mathbf{y})\pi(\mathbf{y}^p))$, state $\mathbf{y}$ becomes state $\mathbf{z}$ while the auxiliary state $\mathbf{y}^p$ remains the same. Alternatively, with probability $\left[1-\dfrac{1}{\pi(\mathbf{y})\pi(\mathbf{y}^p)}\displaystyle\int_{\mathbb{R}^d}\mathfrak{F}(\mathbf{y},\mathbf{y}^p,\mathbf{x},-\mathbf{z}^p)\mathrm{d}\mathbf{x}\right]$, no new state $(\mathbf{x},\mathbf{y}^p)$ is accepted conditioning on currently being at state $(\mathbf{y},\mathbf{y}^p)$. Instead, state $(\mathbf{y},\mathbf{y}^p)$ is directly changed to state $(\mathbf{y},-\mathbf{y}^p)$, leading to a different jump process in $\mathbf{y}$. An illustration of the update rule is shown in Fig. 5.1.

From (2.53), we see that this proposed algorithm takes the anti-symmetric kernel $A(\mathbf{y},\mathbf{y}^p,\mathbf{z},\mathbf{z}^p)$ to be

$$A(\mathbf{y},\mathbf{y}^p,\mathbf{z},\mathbf{z}^p)=\frac{1}{2\Delta t}\Big(\pi(\mathbf{y})\pi(\mathbf{y}^p)p(\mathbf{z},\mathbf{z}^p|\mathbf{y},\mathbf{y}^p;\Delta t)-\pi(\mathbf{z})\pi(\mathbf{z}^p)p(\mathbf{y},\mathbf{y}^p|\mathbf{z},\mathbf{z}^p;\Delta t)\Big) \tag{5.17}$$

with $p(\mathbf{z},\mathbf{z}^p|\mathbf{y},\mathbf{y}^p;\Delta t)$ as in (5.16). To ensure correctness of the sampler, we prove that $A(\mathbf{y},\mathbf{y}^p,\mathbf{z},\mathbf{z}^p)$ must satisfy (condition 3):

$$\int_{\mathbb{R}^{d+1}}A(\mathbf{y},\mathbf{y}^p,\mathbf{z},\mathbf{z}^p)\,\mathrm{d}\mathbf{y}\,\mathrm{d}\mathbf{y}^p=0.$$

Proof 7

$$\begin{aligned}&\int_{\mathbb{R}^{d+dp}}A(\mathbf{x},\mathbf{x}^p,\mathbf{z},\mathbf{z}^p)\,d\mathbf{x}\,d\mathbf{x}^p\\=&\frac{1}{2}\int_{\mathbb{R}^d}\Big(\mathfrak{F}(\mathbf{y},\mathbf{z}^p,\mathbf{z},\mathbf{z}^p)-\mathfrak{F}(\mathbf{z},\mathbf{z}^p,\mathbf{y},\mathbf{z}^p)\Big)\,d\mathbf{y}\\-&\frac{1}{2}\int_{\mathbb{R}^d}\Big(\mathfrak{F}(\mathbf{z},-\mathbf{z}^p,\mathbf{x},-\mathbf{z}^p)-\mathfrak{F}(\mathbf{z},\mathbf{z}^p,\mathbf{x},\mathbf{z}^p)\Big)d\mathbf{x}.\end{aligned}$$

One can check that in (5.19) *and* (5.15), $\widetilde{f}(\mathbf{z},\cdot\,|\mathbf{y},-\mathbf{y}^p)=\widetilde{g}(\mathbf{z},\cdot\,|\mathbf{y},\mathbf{y}^p)$. *Hence,*

$$\mathfrak{F}(\mathbf{y},-\mathbf{y}^p,\mathbf{z},-\mathbf{z}^p)=\mathfrak{F}(\mathbf{z},\mathbf{z}^p,\mathbf{y},\mathbf{y}^p).$$

Therefore

$$\begin{aligned}&\int_{\mathbb{R}^{d+dp}}A(\mathbf{x},\mathbf{x}^p,\mathbf{z},\mathbf{z}^p)\,d\mathbf{x}\,d\mathbf{x}^p \qquad (5.18)\\=&\frac{1}{2}\int_{\mathbb{R}^d}\Big(\mathfrak{F}(\mathbf{z},-\mathbf{z}^p,\mathbf{y},-\mathbf{z}^p)-\mathfrak{F}(\mathbf{z},\mathbf{z}^p,\mathbf{y},\mathbf{z}^p)\Big)\,d\mathbf{y}\\-&\frac{1}{2}\int_{\mathbb{R}^d}\Big(\mathfrak{F}(\mathbf{z},-\mathbf{z}^p,\mathbf{x},-\mathbf{z}^p)-\mathfrak{F}(\mathbf{z},\mathbf{z}^p,\mathbf{x},\mathbf{z}^p)\Big)d\mathbf{x}\\=&0.\end{aligned}$$

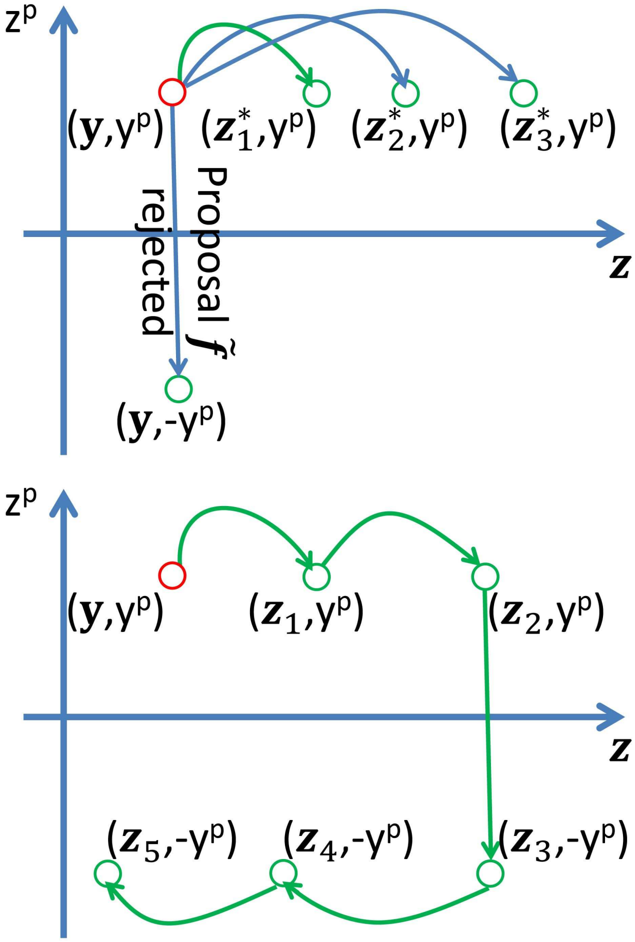


Figure 5.1: Update rule starting from state $(\mathbf{y}, \mathbf{y}^p)$. *Left:* Several possible states $(\mathbf{z}^*, \mathbf{z}^p)$ that the algorithm could visit in the next step. Without resampling the auxiliary variables, $\mathbf{z}^p$ can only be $\mathbf{y}^p$ or $-\mathbf{y}^p$. *Right:* Assuming the algorithm visits $(\mathbf{z}_1, \mathbf{y}^p)$ as the next state to $(\mathbf{y}, \mathbf{y}^p)$ (indicated by the green arrow), a sample trajectory of states generated.

The intuition is that the jump in the auxiliary variable introduces a circulative behavior to the whole process (see Fig. 5.1 for illustration). This circulation of probability flux is exactly balanced with the jumps in the original variable and the auxiliary variable. We also see in Fig. 5.1 that irreversibility introduces a directional effect (just like HMC introduces a direction of rotation). This algorithm is a generalization of the *guided walk Metropolis* method [61] and works well in one dimension, as we demonstrate in Section 5.3.3. In what follows, we generalize this idea to higher dimensions $d > 1$.

Moving to higher dimensions An irreversible sampler in $\mathbb{R}^d$ can be constructed as follows. We expand the state space by introducing a d^p-dimensional auxiliary variable $\mathbf{y}^p \in \mathbb{R}^{d^p}$ in the new state space $(\mathbf{y}, \mathbf{y}^p)$. The total probability can be designated as: $\pi(\mathbf{y}, \mathbf{y}^p) = \pi(\mathbf{y})\pi(\mathbf{y}^p)$. We further impose symmetry on the auxiliary variables such that $\pi(\mathbf{y}^p) = \pi(-\mathbf{y}^p)$, and let

$$\begin{aligned} \widetilde{f}(\mathbf{z}, \mathbf{z}^p|\mathbf{y}, \mathbf{y}^p) &= \prod_{i=1}^{d^p} \big(\mathcal{H}(\mathbf{y}_i^p) f_i(\mathbf{z}|\mathbf{y}, \mathbf{y}_i^p) + \mathcal{H}(-\mathbf{y}_i^p) g_i(\mathbf{z}|\mathbf{y}, -\mathbf{y}_i^p)\big); \\ \widetilde{g}(\mathbf{z}, \mathbf{z}^p|\mathbf{y}, \mathbf{y}^p) &= \prod_{i=1}^{d^p} \big(\mathcal{H}(-\mathbf{y}_i^p) f_i(\mathbf{z}|\mathbf{y}, -\mathbf{y}_i^p) + \mathcal{H}(\mathbf{y}_i^p) g_i(\mathbf{z}|\mathbf{y}, \mathbf{y}_i^p)\big), \end{aligned} \tag{5.19}$$

where $f_i(\mathbf{z}|\mathbf{y}, \mathbf{y}_i^p)$ and $g_i(\mathbf{z}|\mathbf{y}, \mathbf{y}_i^p)$ are conditional probability distributions defined by the value of $\mathbf{y}_i^p$.

This definition of $\widetilde{f}$ and $\widetilde{g}$ is a direct generalization of the definition of (5.15) in the one dimensional case. Fitting this definition into the transition probability $p(\mathbf{z}, \mathbf{z}^p|\mathbf{y}, \mathbf{y}^p; \Delta t)$ in (5.16), the generalized update rule is defined and described in Algorithm 4. Again, we have the anti-symmetric kernel $A(\mathbf{y}, \mathbf{y}^p, \mathbf{z}, \mathbf{z}^p)$ as in (5.17). We can prove—exactly the same as Proof 7—that this construction has $\int_{\mathbb{R}^{d+d^p}} A(\mathbf{y}, \mathbf{y}^p, \mathbf{z}, \mathbf{z}^p)\, \mathrm{d}\mathbf{y}\, \mathrm{d}\mathbf{y}^p = 0$ even with our d^p-dimensional *continuous* auxiliary variables.

In summary, we can use (5.16) to devise a practical algorithm for sampling (Algorithm 4). In particular, if we define $f_i(\mathbf{z}|\mathbf{y}, \mathbf{y}_i^p)$ and $g_i(\mathbf{z}|\mathbf{y}, \mathbf{y}_i^p)$ that are easy to sample from, then we can use the definitions of $\widetilde{f}$ and $\widetilde{g}$ in (5.19) to propose samples in the same way as the MH algorithm. After multiple rejections in $\mathbf{y}$, we resample $\mathbf{y}^p$ according to $\pi(\mathbf{y}^p)$ for a faster-mixing Markov chain in $\mathbf{y}$.

In multiple dimensions, a favorable direction of exploration is often not clear. Hence we suggest to take $d^p = d$ as used in our experiment, so that $\mathbf{z}^p$ has the same dimension as $\mathbf{z}$. Thus all directions can be explored by resampling the auxiliary variable $\mathbf{z}^p$ after multiple rejections. This setting also helps to avoid the possibility of the resulting Markov chain being reducible. Also, when $d^p = d$, f_i and g_i can be designed as: $f_i(\mathbf{z}|\mathbf{y}) = f_i(\mathbf{z}_i|\mathbf{y}_i)$, and $g_i(\mathbf{z}|\mathbf{y}) = g_i(\mathbf{z}_i|\mathbf{y}_i)$, depending only on $\mathbf{z}_i$ and $\mathbf{y}_i$. Sample values in each dimension can thus be independently generated according to $f_i(\mathbf{z}_i|\mathbf{y}_i)$ or $g_i(\mathbf{z}_i|\mathbf{y}_i)$. When a favorable direction of exploration *can* be determined (e.g., in the irreversible MALA algorithm in Section 5.4.3), we can take $d^p = 1$. Then $\mathbf{z}^p$ belongs to a binary set $\{-1, 1\}$, rendering Algorithm 4 the same as the simpler version, Algorithm 3, which is the continuous state space generalization of the *lifting* method [172, 176].

Algorithm 4: Monte Carlo Algorithm from Irreversible Jump Process

```
for $t = 0, 1, 2 \cdots N_{iter}$ do
    optionally, periodically resample auxiliary variable $\mathbf{z}^p$ as $\mathbf{z}^p(t) \sim \pi(\mathbf{z}^p)$
    sample $u \sim \mathcal{U}_{[0,1]}$
    sample $\mathbf{z}(*) \sim \widetilde{f}\left(\mathbf{z}(*), \mathbf{z}^p(*)|\mathbf{z}(t), \mathbf{z}^p(t)\right)$
    $\alpha\left(\mathbf{z}(t), \mathbf{z}^p(t), \mathbf{z}(*), \mathbf{z}^p(*)\right) = \min\left\{1, \frac{\pi\left(\mathbf{z}(*)\right)\pi\left(\mathbf{z}^p(*)\right)\widetilde{g}\left(\mathbf{z}(t), \mathbf{z}^p(t)|\mathbf{z}(*), \mathbf{z}^p(*)\right)}{\pi\left(\mathbf{z}(t)\right)\pi\left(\mathbf{z}^p(t)\right)\widetilde{f}\left(\mathbf{z}(*), \mathbf{z}^p(*)|\mathbf{z}(t), \mathbf{z}^p(t)\right)}\right\}$
    if $u < \alpha\left(\mathbf{z}(t), \mathbf{z}^p(t), \mathbf{z}(*), \mathbf{z}^p(*)\right)$, $\left(\mathbf{z}(t+1), \mathbf{z}^p(t+1)\right) = \left(\mathbf{z}(*), \mathbf{z}^p(t)\right)$
    else $\left(\mathbf{z}(t+1), \mathbf{z}^p(t+1)\right) = \left(\mathbf{z}(t), -\mathbf{z}^p(t)\right)$
end
```

In the experiments of Section 5.3.3, we take $d^p = d$, $f_i\left(\mathbf{z}(*)|\mathbf{z}(t), \mathbf{z}^p(t)\right)$ as $(\mathbf{z}_i(*) - \mathbf{z}_i(t))/\mathbf{z}_i^p(t) \sim \Gamma(\alpha, \beta)$; $g_i\left(\mathbf{z}(*)|\mathbf{z}(t), \mathbf{z}^p(t)\right)$ as $(\mathbf{z}_i(t) - \mathbf{z}_i(*))/\mathbf{z}_i^p(t) \sim \Gamma(\alpha, \beta)$ and let $\pi(\mathbf{z}^p)$ to be a restricted uniform distribution on the set $\left\{\mathbf{z}^p \middle| \frac{1}{N}\left|\mathbf{z}^p\right|_1 = 1\right\}$. Here $\widetilde{f}$ and $\widetilde{g}$ are designed to have no overlap in their support, maximizing the irreversibility effect. The norm of $\mathbf{z}^p$ is set to be constant to ensure that $\mathbf{z}^p$ contributes to the exploration of direction, instead of the expected distance of jump. It is worth noting that the accept-reject step in the current setting is the same as in random-walk MH.

5.3.3 Synthetic Experiments for Irreversible Jump Sampler

To examine the correctness and attributes of our irreversible jump sampler (Algorithm 4), we consider various simulated scenarios, including the challenging cases of heavy tailed, multimodal, and correlated distributions. As mentioned in Section 5.3.2, we take $f_i\left(\mathbf{z}(*)|\mathbf{z}(t), \mathbf{z}^p(t)\right)$ according to $\mathbf{z}_i(*) = \mathbf{z}_i(t) + \gamma\mathbf{z}_i^p(t)$; $g_i\left(\mathbf{z}(*)|\mathbf{z}(t), \mathbf{z}^p(t)\right)$ according to $\mathbf{z}_i(*) = \mathbf{z}_i(t) - \gamma\mathbf{z}_i^p(t)$, where $\gamma \sim \Gamma(\alpha, \beta)$ and let $\pi(\mathbf{z}^p)$ to be a restricted uniform distribution on the set $\left\{\mathbf{z}^p \middle| \frac{1}{N}\left|\mathbf{z}^p\right|_1 = 1\right\}$.

The hyperparameters α and β are chosen using a generic procedure without fine-tuning to the target. We take the gamma shape parameter to be $\alpha = 1.1$, and change the rate parameter β approximately as $\beta \propto \sqrt{V}$ (V is the volume of the region we would like to explore).

1D Heavy-tailed Distribution

We start by considering the task of sampling from 1D normal and log-normal distributions, the latter of which is a heavy-tailed distribution. The motivation for considering the simple 1D normal distribution is to validate the correctness of the sampler and to serve as a comparison relative to the heavy-tailed setting. We compare performance to a MH algorithm with normal proposals centered at the previous state. The results are shown in Fig. 5.2. Some may argue that the main possible benefit of our sampler arises from the gamma proposal distribution. To test this idea, we also compare against an MH algorithm using a symmetrized gamma proposal distribution: $(\mathbf{z}(*) - \mathbf{z}(t)) \sim \frac{1}{2}\left(f\left(\mathbf{z}(*)|\mathbf{z}(t)\right) + g\left(\mathbf{z}(*)|\mathbf{z}(t)\right)\right)$.

We found that the irreversible jump sampler with the gamma proposals has better performance. In particular, the method can decrease autocorrelation without increasing the rejection rate (the rejection rate of all three methods are similar). The MH algorithm with symmetrized gamma proposals, on the other hand, leads to even higher autocorrelation than the vanilla MH algorithm. Intuitively, this result can be understood from Fig. 5.1: the irreversible algorithm leads to further exploration in one direction before circling back. (Also, see Fig. 5.9.)

For the heavy-tailed distribution, similar behavior is observed: the irreversible jump sampler converges to the desired distribution faster because its samples decorrelate more rapidly as a function of run time.

Multimodal Distributions

2D Bimodal distributions We use our irreversible jump sampler to sample increasingly challenging bimodal distributions in 2D, $\pi(z_1, z_2) = 2(z_1^2 - \tau)^2 - 0.2z_1 - 5z_1^2 + 5z_2^2$, displayed in Fig. 5.3. Based on the results of Section 5.3.3, we simply compare against MH with random walk normal proposals and drop the symmetrized gamma proposal case. In Fig. 5.3 we see that the irreversible jump sampler significantly outperforms the random walk MH algorithm. Intuitively, this is facilitated by the greater traversing ability of the irreversible sampler, so that with the same acceptance rate, the irreversible sampler can explore more possible states than the reversible sampler, and have greater chance of transiting into another mode.

One way to capture this difference in the bimodal case is in terms of *escape time* from local modes, which we summarize in Table 5.1. We see that the irreversible jump sampler has escape times orders of magnitude lower. Furthermore, these escape times increase at a much smaller rate as the local modes become more concentrated, indicating much more rapid mixing between modes.

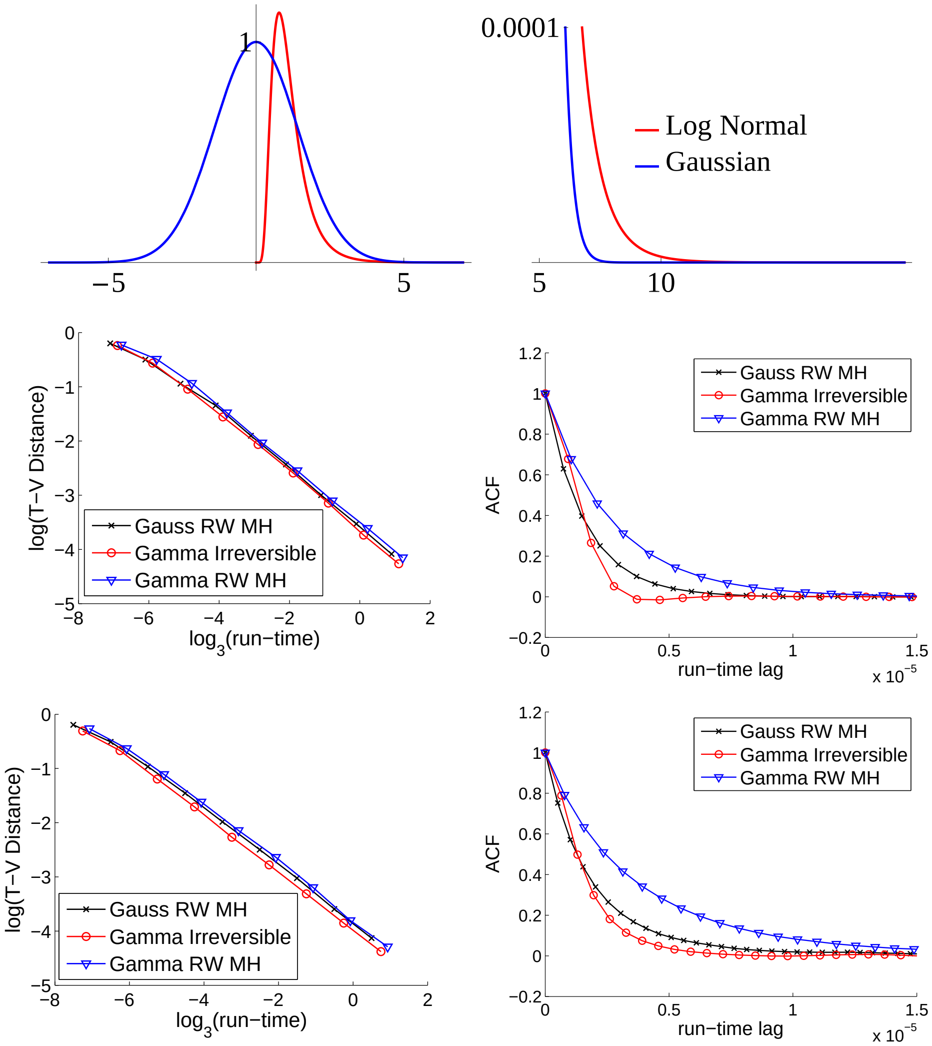


Figure 5.2: *Top row:* (Left) Normal and log-normal target distributions, and (right) zoom in of the tail distributions. *Middle row:* Results for normal target in terms of log total variation distance (T-V distance) vs. log run time (left) and ACF vs. lag in run time (right). *Bottom row:* Analogous plots for log normal target. Comparisons are made among the irreversible jump sampler of Algorithm 4 (Gamma Irreversible), random walk MH algorithm with Gaussian proposals (Gauss RW MH), and random walk MH algorithm with symmetrized gamma proposals (Gamma RW MH). Run time is measured in seconds.

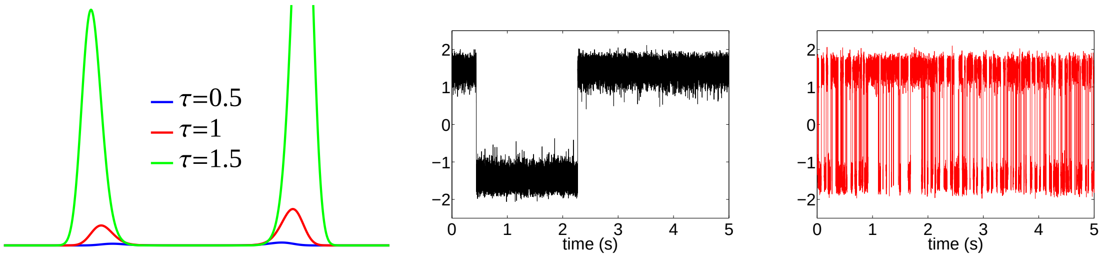


Figure 5.3: (Left) Bimodal targets, $\pi(z_1, z_2) = 2(z_1^2 - \tau)^2 - 0.2z_1 - 5z_1^2 + 5z_2^2$, for various values of τ. Here we demonstrate a 1D cross section of the 2D distribution. (Middle) Sample state trajectories for MH and (right) irreversible jump sampler for the $\tau = 1$ case.

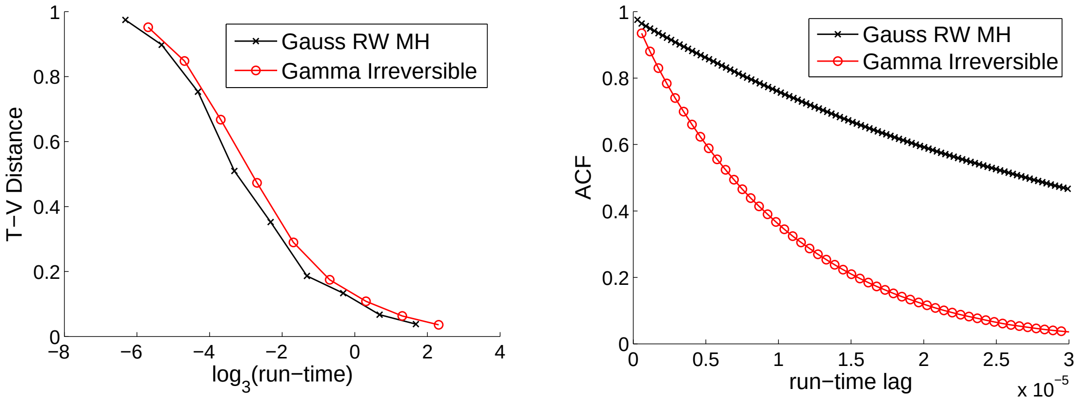


Figure 5.4: Total variational distance vs. log run time (left) and ACF vs. lag in run time (right), for the case of $\tau = 0.5$.

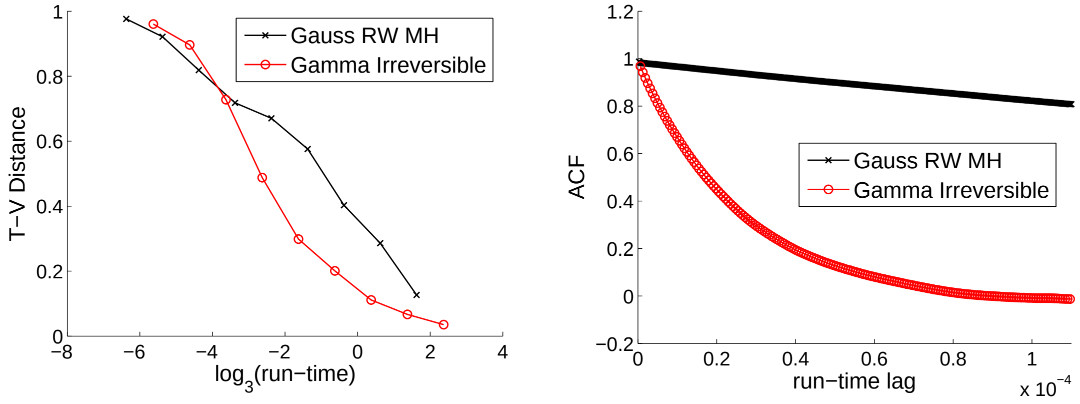


Figure 5.5: Total variational distance vs. log run time (left) and ACF vs. lag in run time (right), for the case of $\tau = 1$.

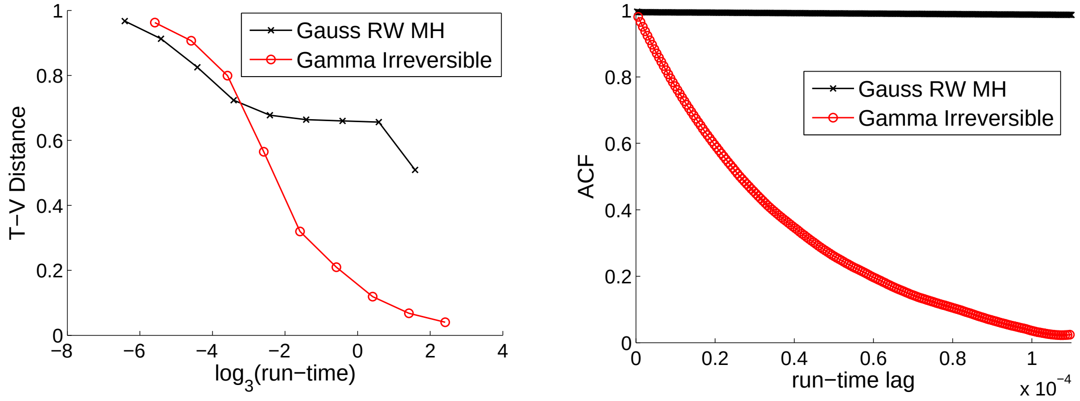


Figure 5.6: Total variational distance vs. log run time (left) and ACF vs. lag in run time (right), for the case of $\tau = 1.5$.

τ	Avg. Escape Time for Irr. Sampler	Avg. Escape Time for MH Sampler
0.5	1.94×10^2	1.06×10^3
1	4.64×10^2	2.47×10^4
1.5	9.06×10^2	7.89×10^5
2	2.41×10^3	N/A

Table 5.1: Comparison of average escape time from one local mode to another between the irreversible jump sampler and random walk MH. The distribution in 2D is more challenging with bigger values of τ (plotted in Fig. 5.3). "N/A" in the last entry means that the escape time is so long that an accurate estimate of it is not available.

2D Multimodal distributions We also tested our method against a recently considered multimodal setting [165]. In the first setting considered, the target distribution is highly multimodal in 2D with unevenly distributed modes. Furthermore, the high mass modes have smaller radii of variation. In the second setting considered, these modes are highly concentrated and well separated, which is an extremely challenging setting for most samplers. See Figs. 5.7 and 5.8. In [165], a *repulsive-attractive Metropolis (RAM)* sampler was proposed with a structure specifically designed to efficiently handle these types of multimodal distributions. We use this as a gold-standard comparison, since this method was already shown to outperform parallel tempering and alternatives [84] in this setting.

We focus our performance analysis on the decay speed of the autocorrelation function (ACF). This can be understood by taking the Gaussian random walk MH algorithm as an example: Although the Gaussian random walk MH algorithm seems to perform well in terms of convergence of total variation distance, this effect is based on exploring one mode really well in a short period of time, instead of making more distant moves to explore other modes. In contrast, the ACF better characterizes the exploration of the samples through the whole space.

Our results are summarized in Figs. 5.7 and 5.8 for each of the two simulated multimodal scenarios. In the first scenario, our sampler outperforms both MH and RAM. In the second scenario, where we have highly concentrated and separated modes, the RAM method tailored to this scenario slightly outperforms our approach. Overall, however, the irreversible jump sampler provides surprisingly good performance in these scenarios despite not having been designed specifically for this setting.

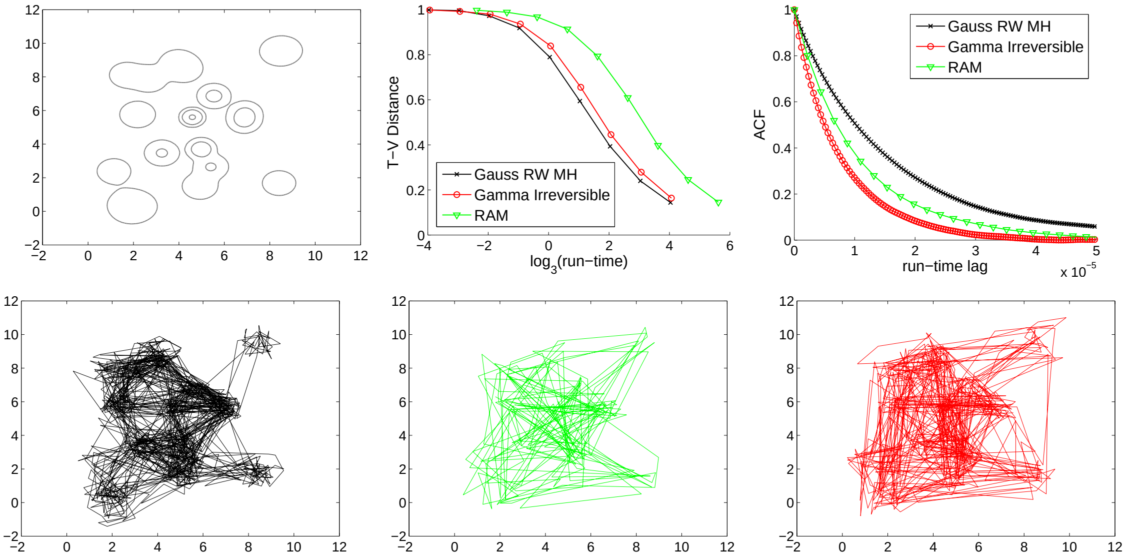


Figure 5.7: *Top row:* (Left) Contour plot of a challenging multimodal probability density function; (middle) T-V distance and ACF comparisons among Gauss RW MH algorithm, Gamma Irreversible, and the recently proposed repulsive-attractive Metropolis (RAM) sampler. *Bottom row:* A sample run of all three samplers, respectively.

5.4 Irreversible Metropolis Adjusted Langevin Algorithm

As discussed in Section 5, there are various ways to combine the continuous dynamics with jump processes to propose new samplers. Since the Markov processes constructed in Section 5.2 and 5.3 can all be non-autonomous (resulting in time dependent matrices $\mathbf{D}$, $\mathbf{Q}$ and kernel functions S, A) as long as the stationary processes converge to the target distribution, one can iteratively follow continuous dynamics and jump processes to propose samples. With ergodicity, averages with respect to the sample values converge to averages with respect to the distribution. This is what is done in HMC: The complete dynamics include (i) a continuous Hamiltonian system with $\mathbf{Q}$ equal to the symplectic matrix, and (ii) a jump process in the auxiliary variable r with the symmetric kernel $S(r_{\mathbf{y}}, r_{\mathbf{z}}) = \pi(r_{\mathbf{y}})\pi(r_{\mathbf{z}})/\Delta t$; the latter corresponds to the resampling of r. Alternating between the two processes provides the HMC method with exploration across the target distribution and ergodicity. (Note that an important consequence of the momentum reversal and resampling, however, is that the resulting HMC dynamics are *reversible*. In contrast, alternating between our proposed

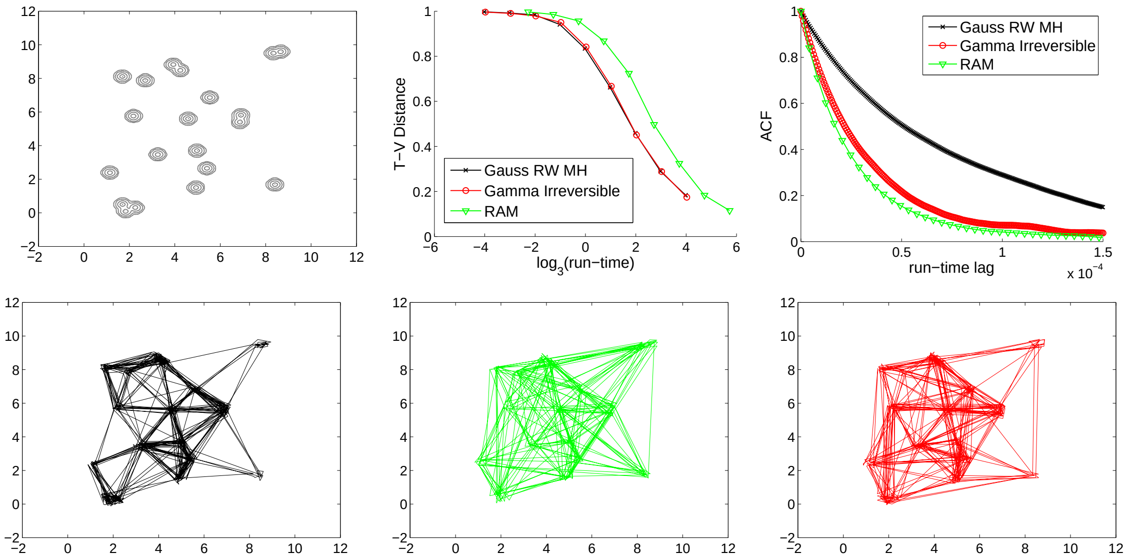


Figure 5.8: Plots as in Fig. 5.7, but for an even more challenging multimodal case where the modes are very concentrated and well separated.

continuous and jump processes can still lead to irreversible dynamics.) Another straightforward way of combining our continuous and jump processes is to use the continuous dynamic sampler for some variables (e.g., real-valued variables) and the jump process sampler for others (e.g., discrete-valued variables).

In addition to the aforementioned means of combining continuous dynamics with jump processes for sampling, in this section we discuss how to use the continuous dynamics as a proposal distribution in our irreversible jump process accept-reject scheme of Algorithm 3, even when the continuous dynamics are not reversible. Previously, similar methods, such as the Metropolis-adjusted Langevin diffusion (MALA) and Riemannian Metropolis-adjusted Langevin diffusion (RMALA) [150, 184, 57], have only been proposed for reversible processes. These methods use one step integration of reversible SDEs to propose samples within a MH algorithm that accepts or rejects the proposal. In this section, we extend these methods to include proposals from any SDE in the form of (5.6) (any SDE with a mild integrability condition), without the requirement of reversibility.

In Sec. 5.4.1, we introduced the MALA algorithm. In Sec. 5.4.2, we discussed how one can use the continuous Markov process for proposal distributions in the irreversible jump sampler and get acceptance rate equal to 1 when the continuous Markov process is simulated exactly. In Sec. 5.4.3, we use a one step simulation of the continuous Markov process for the proposal distribution of the irreversible jump sampler to construct a practical and easy to use algorithm. In Section 5.6, we show that this combination can generate better results in terms of rapid and efficient exploration of a distribution.

5.4.1 Metropolis Adjusted Langevin Algorithm (MALA)

Since MALA algorithm is a special case of the RMALA algorithm (with $\mathbf{D}$ matrix taken to be constant), we will simply introduce RMALA in this section. The RMALA algorithm takes $\mathbf{z} = \theta$ and constructs the proposal distributions $q(\mathbf{z}(*)|\mathbf{z}(t))$ in the MH step (in Algorithm 1) according to the discretized Riemannian Langevin Dynamics:

$$\mathbf{z}(*) \leftarrow \mathbf{z}(t) - \Delta t \cdot [\mathbf{G}(\mathbf{z}(t))^{-1}\nabla U(\mathbf{z}(t)) + \Gamma^{\mathbf{D}}(\mathbf{z}(t))] + \eta(t), \tag{5.20}$$

$$\eta(t) \sim \mathcal{N}(0, 2\Delta t\mathbf{G}(\mathbf{z}(t))^{-1}).$$

Therefore, the transition probability $q(\mathbf{z}(*)|\mathbf{z}(t))$ in the MH Algorithm 1 is:

$$q(\mathbf{z}(*)|\mathbf{z}(t)) \sim \mathcal{N}\big\{\mathbf{z}(*)\big|\mu(\mathbf{z}(t), \Delta t), 2\Delta t\mathbf{G}(\mathbf{z}(t))^{-1}\big\}, \tag{5.21}$$

where

$$\mu(\mathbf{z}(t), \Delta t) = \mathbf{z}(t) - \Delta t \cdot [\mathbf{G}(\mathbf{z}(t))^{-1}\nabla U(\mathbf{z}(t)) + \Gamma^{\mathbf{D}}(\mathbf{z}(t))].$$

This algorithm provides a sampling procedure to exactly simulate the reversible continuous Markov dynamics. And in doing so, gradient information is used to help the sampler efficiently explore the target distribution.

But just as the HMC algorithm, use of the MH procedure inevitably restricts the sampler to be reversible. It can be observed here that only reversible Langevin dynamics are used in the update step of MALA algorithm Eq. (5.20). Although irreversible dynamics can be used in Eq. (5.20), as will be discussed in the Sec. 5.4.2, the acceptance rate would decrease with the increase of irreversibility.

5.4.2 General SDE Proposals under Small Step Size Limit

Our ultimate goal is to use the stochastic dynamics of (5.6) to propose samples in the framework of Algorithm 3. In practice, we need to simulate from the discretized SDE of (5.7). Before analyzing this case, we first examine what would happen if we could *exactly* simulate the SDE of (5.6).

Here, we take $f(\mathbf{z}|\mathbf{y},\mathbf{y}^p)$ in (5.19) to be a Markov transition kernel $P\big(\mathbf{z}|\mathbf{y};\mathrm{d}t\big)$ defined via an infinitesimal step $\mathrm{d}t$ in the SDE:

$$\mathrm{d}\mathbf{z} = \Big[-\big(\mathbf{D}(\mathbf{z})+\mathbf{Q}(\mathbf{z})\big)\nabla H(\mathbf{z}) + \Gamma(\mathbf{z})\Big]\mathrm{d}t + \sqrt{2\mathbf{D}(\mathbf{z})}\mathrm{d}\mathbf{W}(t), \tag{5.22}$$

where $\Gamma_i(\mathbf{z}) = \sum_j \dfrac{\partial}{\partial \mathbf{z}_j}\left(\mathbf{D}_{ij}(\mathbf{z}) + \mathbf{Q}_{ij}(\mathbf{z})\right)$.

For the reverse proposal $g(\mathbf{z}|\mathbf{y},\mathbf{y}^p)$ in (5.19), we use the adjoint process $P^\dagger\big(\mathbf{z}|\mathbf{y};\mathrm{d}t\big)$, inverting the irreversible dynamics via $\mathbf{Q}(\mathbf{z}) \to -\mathbf{Q}(\mathbf{z})$ [105]:

$$\mathrm{d}\mathbf{z} = \Big[-\big(\mathbf{D}(\mathbf{z})-\mathbf{Q}(\mathbf{z})\big)\nabla H(\mathbf{z}) + \widetilde{\Gamma}(\mathbf{z})\Big]\mathrm{d}t + \sqrt{2\mathbf{D}(\mathbf{z})}\mathrm{d}\mathbf{W}(t), \tag{5.23}$$

where $\widetilde{\Gamma}_i(\mathbf{z}) = \sum_j \dfrac{\partial}{\partial \mathbf{z}_j}\left(\mathbf{D}_{ij}(\mathbf{z}) - \mathbf{Q}_{ij}(\mathbf{z})\right)$.

Theorem 2 *For the Markov processes $P\left(\mathbf{z}(T)|\mathbf{z}(t);(T-t)\right)$ and $P^\dagger\left(\mathbf{z}(T)|\mathbf{z}(t);(T-t)\right)$ defined by the SDEs of* (5.22) *and* (5.23) *through Itô integral, the following equality holds:*

$$\frac{P\left(\mathbf{z}(T)|\mathbf{z}(t);(T-t)\right)}{P^\dagger\left(\mathbf{z}(t)|\mathbf{z}(T);(T-t)\right)} = \frac{\pi\left(\mathbf{z}(T)\right)}{\pi\left(\mathbf{z}(t)\right)}. \tag{5.24}$$

Proof 8 *We first prove that for the infinitesimal generators, the backward probability transition kernel following the adjoint process and the forward probability transition kernel are related as:*

$$\pi(\mathbf{y})P\big(\mathbf{z},t+dt|\mathbf{y},t\big) = \pi(\mathbf{z})P^\dagger\big(\mathbf{y},t+dt|\mathbf{z},t\big).$$

Taking path integrals with respect to the infinitesimal generators leads to the conclusion.

As is standard, we use two arbitrary smooth test functions $\psi(\mathbf{y})$ and $\phi(\mathbf{z})$. Then

$$\begin{aligned}
&\int\int d\mathbf{y}d\mathbf{z}\;\psi(\mathbf{y})\phi(\mathbf{z})\;\frac{P\big(\mathbf{z},t+dt|\mathbf{y},t\big)}{\pi(\mathbf{z})}\\
=&\int\int d\mathbf{y}d\mathbf{z}\;\psi(\mathbf{y})\frac{\phi(\mathbf{z})}{\pi(\mathbf{z})}\;\left(P\big(\mathbf{z},t|\mathbf{y},t\big) + \frac{\partial P\big(\mathbf{z},t|\mathbf{y},t\big)}{\partial t}dt\right)\\
=&\int\int d\mathbf{y}d\mathbf{z}\;\psi(\mathbf{y})\frac{\phi(\mathbf{z})}{\pi(\mathbf{z})}\;\left(P\big(\mathbf{z},t|\mathbf{y},t\big) + \mathcal{L}_{\mathbf{z}}\left[\frac{P\big(\mathbf{z},t|\mathbf{y},t\big)}{\pi(\mathbf{z})}\right]dt\right),
\end{aligned}$$

where $\mathcal{L}_{\mathbf{z}}[\varphi(\mathbf{z})] = \nabla^T\cdot\big(\left[\mathbf{D}(\mathbf{z})+\mathbf{Q}(\mathbf{z})\right]\left[\nabla\varphi(\mathbf{z})\pi(\mathbf{z})\right]\big)$ leads to the Fokker-Planck equation of the SDE. It can be checked using (2.45) *and* (2.46) *that $\mathcal{L}^\dagger_{\mathbf{z}}[\varphi(\mathbf{z})] = \mathcal{L}^S_{\mathbf{z}}[\varphi(\mathbf{z})] - \mathcal{L}^A_{\mathbf{z}}[\varphi(\mathbf{z})]$.*

For $P^\dagger\big(\mathbf{y},t+dt|\mathbf{z},t\big)$,

$$\begin{aligned}&\int\!\!\int d\mathbf{y}d\mathbf{z}\;\psi(\mathbf{y})\phi(\mathbf{z})\;\frac{P^\dagger\big(\mathbf{y},t+dt|\mathbf{z},t\big)}{\pi(\mathbf{y})}\\ =&\int\!\!\int d\mathbf{y}d\mathbf{z}\;\frac{\psi(\mathbf{y})}{\pi(\mathbf{y})}\phi(\mathbf{z})\;\left(P^\dagger\big(\mathbf{y},t|\mathbf{z},t\big)+\mathcal{L}^\dagger_{\mathbf{y}}\left[\frac{P^\dagger\big(\mathbf{y},t|\mathbf{z},t\big)}{\pi(\mathbf{y})}\right]dt\right).\end{aligned}$$

Noting that $P\big(\mathbf{z},t|\mathbf{y},t\big)$ and $P^\dagger\big(\mathbf{y},t|\mathbf{z},t\big)$ equal to $\delta(\mathbf{z}-\mathbf{y})$, the zeroth order terms:

$$\int\!\!\int d\mathbf{y}d\mathbf{z}\;\psi(\mathbf{y})\phi(\mathbf{z})\;\frac{P\big(\mathbf{z},t|\mathbf{y},t\big)}{\pi(\mathbf{z})}=\int\!\!\int d\mathbf{y}d\mathbf{z}\;\psi(\mathbf{y})\phi(\mathbf{z})\;\frac{P^\dagger\big(\mathbf{y},t|\mathbf{z},t\big)}{\pi(\mathbf{y})}.$$

Then for the first order terms,

$$\begin{aligned}&\int\!\!\int d\mathbf{y}d\mathbf{z}\;\psi(\mathbf{y})\frac{\phi(\mathbf{z})}{\pi(\mathbf{z})}\;\mathcal{L}_{\mathbf{z}}\left[\frac{P\big(\mathbf{z},t|\mathbf{y},t\big)}{\pi(\mathbf{z})}\right]\\ =&\int\!\!\int d\mathbf{y}d\mathbf{z}\;\psi(\mathbf{y})\;\mathcal{L}^\dagger_{\mathbf{z}}\left[\frac{\phi(\mathbf{z})}{\pi(\mathbf{z})}\right]\;\frac{P\big(\mathbf{z},t|\mathbf{y},t\big)}{\pi(\mathbf{z})}\\ =&\int d\mathbf{z}\;\frac{\psi(\mathbf{z})}{\pi(\mathbf{z})}\;\mathcal{L}^\dagger_{\mathbf{z}}\left[\frac{\phi(\mathbf{z})}{\pi(\mathbf{z})}\right]\\ =&\int d\mathbf{z}\;\mathcal{L}_{\mathbf{z}}\left[\frac{\psi(\mathbf{z})}{\pi(\mathbf{z})}\right]\;\frac{\phi(\mathbf{z})}{\pi(\mathbf{z})}\\ =&\int\!\!\int d\mathbf{y}d\mathbf{z}\;\mathcal{L}_{\mathbf{y}}\left[\frac{\psi(\mathbf{y})}{\pi(\mathbf{y})}\right]\;\phi(\mathbf{z})\;\frac{P^\dagger\big(\mathbf{y},t|\mathbf{z},t\big)}{\pi(\mathbf{y})}\\ =&\int\!\!\int d\mathbf{y}d\mathbf{z}\;\frac{\psi(\mathbf{y})}{\pi(\mathbf{y})}\;\phi(\mathbf{z})\;\mathcal{L}^\dagger_{\mathbf{y}}\left[\frac{P^\dagger\big(\mathbf{y},t|\mathbf{z},t\big)}{\pi(\mathbf{y})}\right].\end{aligned}$$

Hence,

$$\begin{aligned}&\int\!\!\int d\mathbf{y}d\mathbf{z}\;\psi(\mathbf{y})\phi(\mathbf{z})\;\frac{P\big(\mathbf{z},t+dt|\mathbf{y},t\big)}{\pi(\mathbf{z})}\\ =&\int\!\!\int d\mathbf{y}d\mathbf{z}\;\psi(\mathbf{y})\frac{\phi(\mathbf{z})}{\pi(\mathbf{z})}\;\left(P\big(\mathbf{z},t|\mathbf{y},t\big)+\mathcal{L}_{\mathbf{z}}\left[\frac{P\big(\mathbf{z},t|\mathbf{y},t\big)}{\pi(\mathbf{z})}\right]dt\right)\\ =&\int\!\!\int d\mathbf{y}d\mathbf{z}\;\frac{\psi(\mathbf{y})}{\pi(\mathbf{y})}\phi(\mathbf{z})\;\left(P^\dagger\big(\mathbf{y},t|\mathbf{z},t\big)+\mathcal{L}^\dagger_{\mathbf{y}}\left[\frac{P^\dagger\big(\mathbf{y},t|\mathbf{z},t\big)}{\pi(\mathbf{y})}\right]dt\right)\\ =&\int\!\!\int d\mathbf{y}d\mathbf{z}\;\psi(\mathbf{y})\phi(\mathbf{z})\;\frac{P^\dagger\big(\mathbf{y},t+dt|\mathbf{z},t\big)}{\pi(\mathbf{y})}.\end{aligned}$$

Therefore, to the first order,

$$\pi(\mathbf{y})P\big(\mathbf{z}, t + dt|\mathbf{y}, t\big) = \pi(\mathbf{z})P^{\dagger}\big(\mathbf{y}, t + dt|\mathbf{z}, t\big).$$

Using the Markov properties,

$$P(\mathbf{z}_N, t_N|\mathbf{z}_0, t_0) = \int \cdots \int \prod_{i=1}^{N-1} d\mathbf{z}_i \prod_{i=0}^{N-1} P(\mathbf{z}_{i+1}, t_{i+1}|\mathbf{z}_i, t_i);$$

and

$$\begin{aligned} P^{\dagger}(\mathbf{z}_0, t_N|\mathbf{z}_N, t_0) =& \int \cdots \int \prod_{i=1}^{N-1} d\mathbf{z}_i \prod_{i=0}^{N-1} P^{\dagger}(\mathbf{z}_i, t_{i+1}|\mathbf{z}_{i+1}, t_i) \\ =& \int \cdots \int \prod_{i=1}^{N-1} d\mathbf{z}_i \prod_{i=0}^{N-1} \frac{\pi(\mathbf{z}_i)}{\pi(\mathbf{z}_{i+1})} P(\mathbf{z}_{i+1}, t_{i+1}|\mathbf{z}_i, t_i) \\ =& \int \cdots \int \prod_{i=1}^{N-1} d\mathbf{z}_i \prod_{i=0}^{N-1} \frac{\pi(\mathbf{z}_i)}{\pi(\mathbf{z}_{i+1})} P(\mathbf{z}_{i+1}, t_{i+1}|\mathbf{z}_i, t_i) \\ =& \frac{\pi(\mathbf{z}_0)}{\pi(\mathbf{z}_N)} P(\mathbf{z}_N, t_N|\mathbf{z}_0, t_0). \end{aligned}$$

Taking the time interval between t_i and t_{i+1} to be infinitesimal, we obtain that

$$\frac{P(\mathbf{z}^{(T)}, T|\mathbf{z}^{(t)}, t)}{P^{\dagger}(\mathbf{z}^{(t)}, T|\mathbf{z}^{(T)}, t)} = \frac{\pi(\mathbf{z}^{(T)})}{\pi(\mathbf{z}^{(t)})}.$$

Analysis on the semigroups $e^{t\mathcal{L}}$ and $e^{t\mathcal{L}^{\dagger}}$ generated by $\mathcal{L}$ and $\mathcal{L}^{\dagger}$ can also lead to this conclusion.

Using Theorem 2, we have

$$\alpha\left(\mathbf{z}(t), \mathbf{z}(*)\right) = \min\left\{1, \frac{\pi\left(\mathbf{z}(*)\right) P^{\dagger}\left(\mathbf{z}(t)|\mathbf{z}(*)\right)}{\pi\left(\mathbf{z}(t)\right) P\left(\mathbf{z}(*)|\mathbf{z}(t)\right)}\right\} = 1. \tag{5.25}$$

Even though in Section 5.2 we saw that SDEs of the form in (5.6) have $\pi(\mathbf{z})$ as the invariant distribution, it is not immediately obvious that using this SDE as a proposal in Algorithm 3 would lead to an acceptance rate of 1. In fact, if we simply use the forward (or backward) transition kernel $P(\mathbf{z}|\mathbf{y}; \mathrm{d}t)$ in the MH algorithm, then the acceptance rate:

$$\alpha^{MH}\left(\mathbf{z}(t), \mathbf{z}(*)\right) = \min\left\{1, \frac{\pi\left(\mathbf{z}(*)\right) P\left(\mathbf{z}(t)|\mathbf{z}(*)\right)}{\pi\left(\mathbf{z}(t)\right) P\left(\mathbf{z}(*)|\mathbf{z}(t)\right)}\right\} \neq 1.$$

And the more irreversibility is introduced, the less the acceptance rate α^{MH} will be in the MH algorithm. This gap of α^{MH} to 1 has been first discovered in statistical mechanics literatures and related to the "house keeping heat" by Crooks and Hatano and Sasa [33, 64, 105] (See Sec. 3.2.4 for a detailed discussion).

Eq. (5.25) also gives us insight into the fact that using more accurate numerical integrators could lead to higher acceptance rates. In Section 5.4.3, we analyze the accept-reject scheme for the simple first-order integration of (5.7) with finite step size Δt.

5.4.3 Irreversible MALA via Irreversible Jump Correction

Since in practice we rely on finite step sizes $\Delta t > 0$, there will be numerical error and $\frac{P\left(\mathbf{z}(*)|\mathbf{z}(t);\Delta t\right)}{P^{\dagger}\left(\mathbf{z}(t)|\mathbf{z}(*);\Delta t\right)}$ can differ from $\frac{\pi\left(\mathbf{z}(*)\right)}{\pi\left(\mathbf{z}(t)\right)}$. We now propose an irreversible generalization of the MALA algorithm to correct for these errors. We make use of Algorithm 3 and take a general SDE and its adjoint process defined in Section 5.4.2 to propose samples using a one-step numerical integration (as in MALA). Because we have the local gradient information in the SDEs to guide us, the direction of the exploration is determined. So, we simply use a 1-dimensional discrete auxiliary variable $\mathbf{y}^p$, and thus the use of Algorithm 3 instead of the more general Algorithm 4. We call the resulting algorithm the *irreversible MALA* method.

Assuming a one-step numerical integration uses a Δt period of time, then the discretization of the SDE of (5.22) leads to

$$P(\mathbf{z}|\mathbf{y};\Delta t) \sim \mathcal{N}\big\{\mathbf{z}|\mu(\mathbf{y},\Delta t), 2\Delta t \cdot \mathbf{D}(\mathbf{y})\big\}, \tag{5.26}$$

where

$$\mu(\mathbf{y},\Delta t) = \mathbf{y} + \Big[-\big(\mathbf{D}(\mathbf{y})+\mathbf{Q}(\mathbf{y})\big)\nabla H(\mathbf{y}) + \Gamma(\mathbf{y})\Big]\Delta t, \Gamma_i(\mathbf{z}) = \sum_j \frac{\partial}{\partial \mathbf{z}_j}\left(\mathbf{D}_{ij}(\mathbf{z}) + \mathbf{Q}_{ij}(\mathbf{z})\right).$$

Importantly, this allows us to compute $f\left(\mathbf{z}(*)|\mathbf{z}(t)\right) = P(\mathbf{z}(*)|\mathbf{z}(t);\Delta t)$ in Algorithm 3. The corresponding calculation for the adjoint process with the SDE in (5.23) is:

$$P^{\dagger}(\mathbf{z}|\mathbf{y};\Delta t) \sim \mathcal{N}\big\{\mathbf{z}|\mu^{\dagger}(\mathbf{y},\Delta t), 2\Delta t \cdot \mathbf{D}(\mathbf{y})\big\}, \tag{5.27}$$

where

$$\mu^{\dagger}(\mathbf{y},\Delta t) = \mathbf{y} + \Big[-\big(\mathbf{D}(\mathbf{y})-\mathbf{Q}(\mathbf{y})\big)\nabla H(\mathbf{y}) + \widetilde{\Gamma}(\mathbf{y})\Big]\Delta t, \widetilde{\Gamma}_i(\mathbf{z}) = \sum_j \frac{\partial}{\partial \mathbf{z}_j}\left(\mathbf{D}_{ij}(\mathbf{z}) - \mathbf{Q}_{ij}(\mathbf{z})\right).$$

This allows us to compute $g\left(\mathbf{z}(*)|\mathbf{z}(t)\right) = P^{\dagger}(\mathbf{z}(*)|\mathbf{z}(t);\Delta t)$. The resulting irreversible MALA algorithm is summarized in Algorithm 5.

Algorithm 5: Irreversible MALA

```
randomly pick $z^p$ from $\{1, -1\}$ with equal probability
for $t = 0, 1, 2 \cdots N_{iter}$ do
    optionally, periodically resample auxiliary variable $z^p \sim \mathcal{U}_{\{1,-1\}}$
    sample $u \sim \mathcal{U}_{[0,1]}$
    if $z^p > 0$ then
        sample $\eta_t \sim \mathcal{N}(0, 2\epsilon_t \mathbf{D}(\mathbf{z}_t))$
        $\mathbf{z}(*) \leftarrow \mathbf{z}_t - \epsilon_t \left[\big(\mathbf{D}(\mathbf{z}_t) + \mathbf{Q}(\mathbf{z}_t)\big)\nabla H(\mathbf{z}_t) + \Gamma(\mathbf{z}_t)\right] + \eta_t$
        $\alpha\left(\mathbf{z}(t), \mathbf{z}(*)\right) = \min\left\{1, \frac{\pi\left(\mathbf{z}(*)\right) P^{\dagger}(\mathbf{z}(t)|\mathbf{z}(*); \Delta t)}{\pi\left(\mathbf{z}(t)\right) P(\mathbf{z}(*)|\mathbf{z}(t); \Delta t)}\right\}$
    end
    else
        sample $\eta_t \sim \mathcal{N}(0, 2\epsilon_t \mathbf{D}(\mathbf{z}_t))$
        $\mathbf{z}(*) \leftarrow \mathbf{z}_t - \epsilon_t \left[\big(\mathbf{D}(\mathbf{z}_t) - \mathbf{Q}(\mathbf{z}_t)\big)\nabla H(\mathbf{z}_t) + \widetilde{\Gamma}(\mathbf{z}_t)\right] + \eta_t$
        $\alpha\left(\mathbf{z}(t), \mathbf{z}(*)\right) = \min\left\{1, \frac{\pi\left(\mathbf{z}(*)\right) P(\mathbf{z}(t)|\mathbf{z}(*); \Delta t)}{\pi\left(\mathbf{z}(t)\right) P^{\dagger}(\mathbf{z}(*)|\mathbf{z}(t); \Delta t)}\right\}$
    end
    if $u < \alpha\left(\mathbf{z}(t), \mathbf{z}(*)\right)$, $\mathbf{z}(t+1) = \mathbf{z}(*)$; $z^p(t+1) = z^p(t)$
    else $\mathbf{z}(t+1) = \mathbf{z}(t)$; $z^p(t+1) = -z^p(t)$
end
```

We know from Section 5.4.2 that in the small Δt limit,

$$\alpha\left(\mathbf{z}(t), \mathbf{z}(*)\right) = \min\left\{1, \frac{P^{\dagger}\left(\mathbf{z}(t)|\mathbf{z}(*)\right)}{P\left(\mathbf{z}(*)|\mathbf{z}(t)\right)} \cdot \frac{\pi\left(\mathbf{z}(*)\right)}{\pi\left(\mathbf{z}(t)\right)}\right\} \to 1. \tag{5.28}$$

From this result, we see that there seems to be a step-size/acceptance-rate tradeoff. As mentioned in Section 5.4.2, a higher-order numerical scheme could potentially increase the acceptance rate with the same step size [26, 20, 94]. We leave this as a direction for future research.

5.5 Related Work

There have been previous efforts to construct irreversible Markov processes for sampling. One example is using continuous dynamics to achieve this goal, which has been studied extensively. One can make use of Hamiltonian or generalized Hamiltonian dynamics to introduce irreversibility into the sampling procedure [70, 71, 145, 39, 119]. There have been other samplers that utilize irreversible continuous dynamics such as underdamped Langevin [68, 28] and Nosé-Hoover [37, 161] dynamics, although irreversibility was not the emphasis in these works. As described in Section 5.2, any dynamic process that has a nonzero $\mathbf{Q}$ matrix can be used to devise an irreversible sampler within our framework. As mentioned in Sec. 6.1.1, the problem is that simulating the continuous Markov processes using the discretized system typically leads to the introduction of bias due to discretization error. And if MH procedure is introduced, the whole process becomes reversible again.

We have also discussed using jump processes for sampling tasks. However, only recently have researchers constructed irreversible jump processes that form valid sampling procedures. In the *non-reversible MH* algorithm [15], a vorticity function (or matrix) is added to the MH procedure. Hence, the difficulty of construction is translated to defining a valid vorticity function, similar to the difficulty of defining the antisymmetric kernel $A(\mathbf{y}, \mathbf{z})$. For the multivariate Gaussian distribution, the author discretized an irreversible Ornstein-Uhlenbeck process to obtain a suitable vorticity function. The *lifting method* [172, 176] makes a replica of the original state space ($\mathbb{R}^d \times \{-1, 1\}$) to facilitate irreversibility in the sampling procedure. A skew detailed balance condition is imposed to ensure a valid anti-symmetric kernel $A(\mathbf{y}, \mathbf{z})$ in the expanded state space. The authors showed an example of applying the method to spin models. For both the *non-reversible MH* and *lifting* methods, it has not been clear how to come up with a practical, easy-to-construct algorithm to handle a broad set of target distributions. Our irreversible jump sampler can be seen as a combination of both ideas which generalizes to arbitrary target distributions. The ideas of lifting the state space (to $\mathbb{R}^d \times \mathbb{R}^{d^p}$) and using an irreversible accept-reject procedure similar to the non-reversible MH algorithm are incorporated into one simple procedure.

Recently, the combined approach of using both continuous dynamics and jump processes has been proposed for constructing irreversible samplers. The *bouncy particle* [21] and *Zig-Zag* [17, 16] samplers use deterministic dynamics (irreversible in nature) combined with a Poisson process to create valid MCMC procedures. These two methods alternate between continuous dynamics and a Poisson jump process with an inhomogeneous rate (or intensity) to ensure the invariance of the target distribution. Our irreversible MALA algorithm avoids the difficulty of sampling from a Poisson process. Additionally, we end up

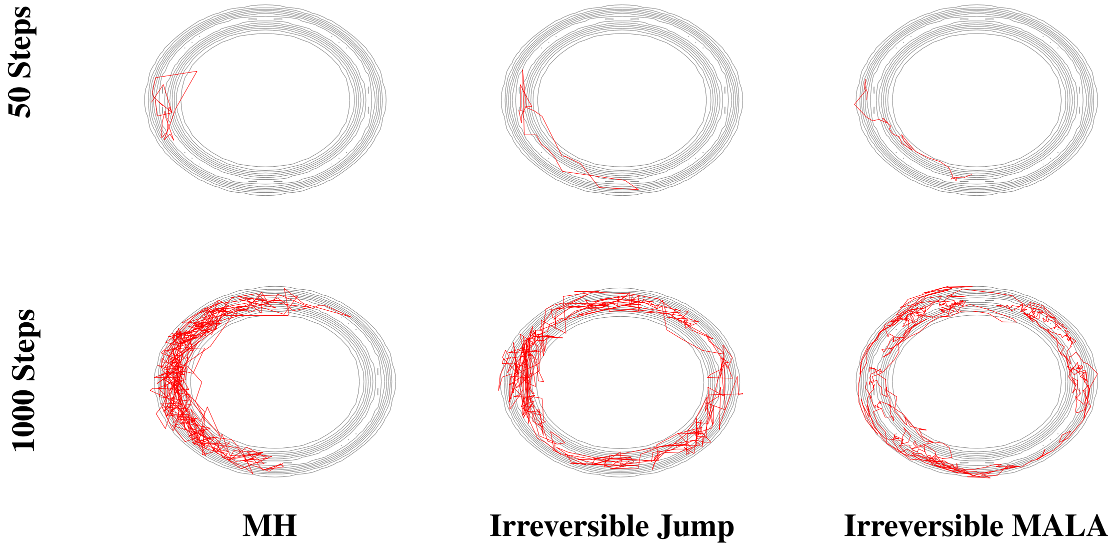


Figure 5.9: *Top row:* Trajectory of first 50 steps of (left) MH algorithm using Gaussian random walk proposals, (middle) irreversible jump algorithm with gamma proposals and (right) irreversible MALA algorithm. *Bottom row:* Similarly for the first 1000 steps of the algorithms.

with an algorithm that is a simple modification of vanilla MH, making it straightforward to use and plug in to existing algorithmic frameworks.

5.6 Experiments

In this section, we explore the irreversible MALA algorithm and compare it against the MH, irreversible jump, MALA, and HMC.

5.6.1 Visual Comparison of Samplers

We first perform a qualitative comparison between the MH algorithm, our irreversible jump sampler (with gamma proposals), and the irreversible MALA algorithm to provide insights into their differences. It is demonstrated in Fig. 5.9 that the standard MH sampler jumps around randomly, but does so within a local region of the previous sample and irrespective of previous (directions of) jumps, leading to slow exploration of the distribution. In contrast, our irreversible counterpart (using gamma proposals) more rapidly traverses the distribution by following the direction of the previous jump, until being rejected. Finally, the irreversible MALA algorithm provides an even smoother trajectory by using continuous dynamics in place of independent gamma proposals.

Having visually examined the differences between the samplers to gain intuition, in what follows we provide a more quantitative analysis of convergence speed in the case of a correlated synthetic distribution. We then compare different methods in a Bayesian logistic regression model and a stochastic volatility model.

5.6.2 Synthetic Example

We also test the correctness and attributes of our algorithm on a highly correlated (moon-shaped) target distribution, where $\pi(z_1, z_2) = z_1^4/10 + (4(z_2 + 1.2) - z_1^2)^2/2$. In terms of number of iterations, the irreversible jump sampler with gamma proposals decorrelates and converges to the posterior distribution faster. However, in terms of run time, irreversible jump sampler does not perform as well as random walk MH algorithm, as explored in Fig. 5.11. The reason is that the correlated distribution has complex geometry. Faster exploration in random directions, as provided by our irreversible sampler with independent proposals, only marginally increases the mixing effect in each step relative to the reversible independent proposals of MH. Since the calculation of the distribution is not demanding in this case, the small overhead of the irreversible sampler (keeping track of the number of rejections and resample the direction of exploration after multiple rejections) actually makes a difference and thus results in our sampler with gamma proposals providing slightly worse performance in terms of runtime.

To improve the performance of our irreversible sampler further in this correlated target case, it would be appealing to take the geometric information about the level sets—including the higher mass regions—into account. Indeed, we are able to do this by replacing the independent gamma proposals with proposals from our continuous dynamics sampler, as described in Section 5.4. To demonstrate the effect of irreversibility, we choose $\mathbf{D}(\mathbf{z}) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ and $\mathbf{Q}(\mathbf{z}) = \begin{pmatrix} 0 & -2 \\ 2 & 0 \end{pmatrix}$ in Eqs. (5.22), (5.23), (5.26), and (5.27). In this case, our irreversible MALA algorithm (Algorithm 5) significantly outperforms the Gaussian random walk MH, as well as HMC [113] and the standard reversible MALA algorithm [150]. Because the target distribution has complex geometry, the continuous dynamics can provide guidance on locating the higher mass regions and exploring the contours rapidly with the gradient information. HMC and MALA algorithms exploit this effect, but we additionally see gains from the irreversibility of the sampler.

This experiment demonstrates the gains that are possible by combining our continuous dynamics and jump process frameworks, beyond what either can provide individually.

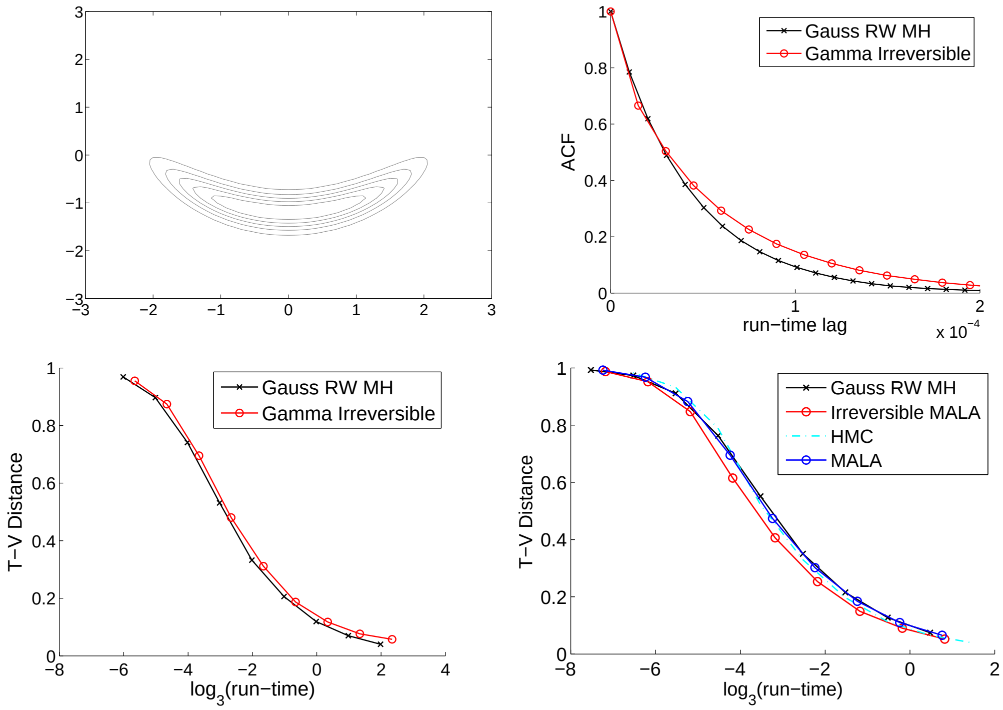


Figure 5.10: *Top row:* Correlated distribution with complex geometry in 2D, $\pi(z_1, z_2) = z_1^4/10 + (4(z_2 + 1.2) - z_1^2)^2/2$ (left) and ACF vs. lag in run time of Gamma Irreversible algorithm against Gauss RW MH (right). *Bottom row:* T-V distance vs. log run time. Comparisons are made between Gauss RW MH and Gamma Irreversible (left), and Gauss RW MH, Irreversible MALA, HMC, and MALA (right).

5.6.3 Bayesian Logistic Regression

In this section, we demonstrate results from sampling a Bayesian logistic regression model. Similar to the setting in [57], we consider an $N \times D$ design matrix $\mathbf{X}$ comprising N samples each with D covariates and a binary response variable $\mathbf{t} \in \{0, 1\}^N$. If we denote the logistic link function by $s(\cdot)$, a Bayesian logistic regression model of the binary response [55, 99] is obtained by the introduction of regression coefficients $\beta \in \mathbb{R}^D$ with an appropriate prior, which for illustration is given as $\beta \sim \mathcal{N}(\mathbf{0}, \alpha\mathbf{I})$ where α is given.

We make use of three data sets available at the STATLOG project[1]. The first two are

[1] Link to the project can be found at: https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/

Data Set \ Methods	MH	Irreversible Jump	MALA	Irreversible MALA	HMC
Australian Credit	4.30	10.51	9.05	15.95	10.96
German Credit	2.82	2.88	3.67	4.47	3.73
Heart	22.92	24.14	29.26	41.23	30.83

Table 5.2: Comparison of ESS/sec of the samplers.

datasets describing the connections between credit card approval and various attributes of the applicants in Australian and German. The third dataset is about the connection between the absence or presence of heart disease and various patient information.

Performance of the samplers is measured by the per second effective sample size (ESS/runtime), where ESS [99, 149, 8, 111] is calculated through number of steps N divided by the integrated autocorrelation time τ_{int}: $ESS = \frac{N}{\tau_{int}}$. In [57], τ_{int} is estimated through the initial positive sequence estimator [56]: $\tau_{int} = 1 + 2\sum_k \gamma(k)$, where $\gamma(k)$ is the k-lagged autocorrelations and sum is over the K monotone sample autocorrelations. The initial positive sequence estimator possesses better consistency but assumes the Markov chain to be reversible. We are concerned that this may underestimate the integrated autocorrelation time for irreversible chains. Hence we use the original window estimator [56] for the integrated autocorrelation time τ_{int}, so that:

$$ESS = \frac{N}{\tau_{int}} = \frac{N}{1 + 2\sum_{k=1}^{M}\left(1 - \frac{k}{M}\right)\gamma(k)}, \tag{5.29}$$

where M is a large number (taken to be 3000 in the experiments).

Optimal hyperparameters are found for the methods from a grid search. For MH, we corroborate that the optimal hyperparameters are indeed obtained by tuning the acceptance rate between 20% and 40% at stationary. For HMC, we found that using 10 leap-frog steps to generate a sample is most efficient in terms of ESS/runtime (as opposed to the commonly used 50 to 100 steps). For MALA, acceptance rate between 40% and 60% generates best ESS/runtime. For the irreversible MALA algorithm, we simply take $\mathbf{D} = \mathbf{I}$, and $\mathbf{Q}(\mathbf{z}) = \begin{pmatrix} 0 & -\mathbf{I}_{d\times d} \\ \mathbf{I}_{d\times d} & 0 \end{pmatrix}$ for the first $d = \lfloor D/2 \rfloor$ variables.

5.7 Scaling Up the Sampling Algorithms for Large Datasets

We wish to scale up the previously discussed sampling algorithms to cases where our target distribution is a posterior distribution in a Bayesian model, and we are faced with a huge number of observations $\mathcal{S}$. In this case, the likelihood, or its gradient, can be computationally prohibitive to compute. In the cases of i.i.d. data, we write our target distribution as $\pi(\theta) = p(\mathcal{S}|\theta)\ p(\theta) = \prod_{\mathbf{s}\in\mathcal{S}} p(\mathbf{s}|\theta)\ p(\theta)$. For the samplers designed from continuous dynamics (Section 5.2), we can use *stochastic gradients*, in place of the full data gradient as elaborated in [102] and outlined below. For samplers using jump processes (Section 2.4), we discuss a generalization of the subsampling-within-MH ideas in [82, 11, 12].

Stochastic Gradient Samplers For samplers using continuous dynamics (5.6), the computationally intensive component in the update rule of (5.7) is the computation of $\nabla H(\theta) = -\nabla \log \pi(\theta) = -\sum_{\mathbf{s}\in\mathcal{S}} \nabla \log p(\mathbf{s}|\theta) - \nabla \log p(\theta)$. One idea for avoiding this per iteration cost is to use *stochastic gradients* instead [148]. Here, a noisy gradient based on a data subsample or *minibatch*, is used as an unbiased estimator of the full data gradient. More formally, we examine *independently sampled* minibatches $\widetilde{\mathcal{S}} \subset \mathcal{S}$. The corresponding log-posterior for these data is

$$\widetilde{H}(\theta) = -\frac{|\mathcal{S}|}{|\widetilde{\mathcal{S}}|} \sum_{\mathbf{s}\in\widetilde{\mathcal{S}}} \log p(\mathbf{s}|\theta) - \log p(\theta); \quad \widetilde{\mathcal{S}} \subset \mathcal{S}. \tag{5.30}$$

The specific form of (6.22) implies that $\widetilde{H}(\theta)$ is an unbiased estimator of $H(\theta)$, thus $\nabla\widetilde{H}(\theta)$ is an unbiased estimator of $\nabla H(\theta)$. The key question in many of the existing stochastic gradient MCMC algorithms is whether the noise injected by the stochastic gradient $\nabla\widetilde{H}(\theta)$ adversely affects the stationary distribution of the modified dynamics. One way to analyze the impact of the stochastic gradient is to assume the central limit theorem holds: $\nabla\widetilde{H}(\theta) = \nabla H(\theta) + \mathcal{N}(0, \mathbf{V}(\theta))$. Simply plugging in $\nabla\widetilde{H}(\theta)$ in place of $\nabla H(\theta)$ in (5.7) results in dynamics with an additional noise term $\big(\mathbf{D}(\theta_t) + \mathbf{Q}(\theta_t)\big)[\mathcal{N}(0, \mathbf{V}(\theta_t))]^T$. In our earlier work [102], we studied the influence of this noise and showed that one may counteract it by assuming an estimate $\hat{\mathbf{B}}_t$ of the noise variance and following:

$$\mathbf{z}_{t+1} \leftarrow \mathbf{z}_t - \epsilon_t \Big[\big(\mathbf{D}(\mathbf{z}_t) + \mathbf{Q}(\mathbf{z}_t)\big)\nabla\widetilde{H}(\mathbf{z}_t) + \Gamma(\mathbf{z}_t)\Big] + \mathcal{N}(0, \epsilon_t(2\mathbf{D}(\mathbf{z}_t) - \epsilon_t\hat{\mathbf{B}}_t)). \tag{5.31}$$

In the limit of ϵ_t going to zero, the stationary distribution is preserved. For finite ϵ_t, a bias exists. The same bias-speed tradeoff was used in past stochastic gradient sampling methods [182, 28, 37, 161]. In [102], we also devise methods for defining new samplers using existing $\mathbf{D}$ and $\mathbf{Q}$ matrices as building blocks. We will discuss this topic in more depth in Sec. 6.1.

Subsampling of Irreversible Sampler from Jump Processes For the irreversible jump sampler (Algorithm 4), we can directly generalize the subsampling idea for the MH algorithms [82] and its adaptive and proxy method variations [11, 12]. In Algorithm 4, the computational bottleneck is at the step where we decide to accept or reject the proposal from $\widetilde{f}\,(\theta(*), \theta^p(*)|\theta(t), \theta^p(t))$, since we need to calculate $\pi(\theta(*))/\pi(\theta(t))$, requiring evaluation of the entire likelihood.

The accept-reject step is then implemented by sampling a uniform random variable $u \sim \mathcal{U}_{[0,1]}$ and accepting the proposal if and only if $u < \alpha\,(\theta(t), \theta^p(t), \theta(*), \theta^p(*))$. Following [82, 11], we can rewrite this condition as:

$$\Lambda_{\mathcal{S}}(\theta(t), \theta(*)) > \Theta_{\mathcal{S}}(u, \theta(t), \theta(*), \theta^p(t), \theta^p(*)), \tag{5.32}$$

where

$$\Lambda_{\mathcal{S}}(\theta(t), \theta(*)) = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{s} \in \mathcal{S}} \log \left[\frac{p(\mathbf{s}|\theta(*))}{p(\mathbf{s}|\theta(t))} \right],$$

$$\Theta_{\mathcal{S}}(u, \theta(t), \theta(*), \theta^p(t), \theta^p(*)) = \frac{1}{|\mathcal{S}|} \log \left[u \frac{\pi\,(\theta^p(t))\, \widetilde{f}\,(\theta(*), \theta^p(*)|\theta(t), \theta^p(t))}{\pi\,(\theta^p(*))\, \widetilde{g}\,(\theta(t), \theta^p(t)|\theta(*), \theta^p(*))} \right].$$

For the computationally intractable $\Lambda_{\mathcal{S}}(\theta(t), \theta(*))$, we can use a subset of data to approximate it via:

$$\Lambda^*_{\widetilde{\mathcal{S}}}(\theta(t), \theta(*)) = \frac{1}{|\widetilde{\mathcal{S}}|} \sum_{\mathbf{s} \in \widetilde{\mathcal{S}}} \log \left[\frac{p(\mathbf{s}|\theta(*))}{p(\mathbf{s}|\theta(t))} \right] \approx \Lambda_{\mathcal{S}}(\theta(t), \theta(*)); \quad \widetilde{\mathcal{S}} \subset \mathcal{S}.$$

Importantly, $|\Lambda_{\mathcal{S}}(\theta(t), \theta(*)) - \Lambda^*_{\widetilde{\mathcal{S}}}(\theta(t), \theta(*))|$ can be bounded probabilistically [11]. Hence, a speed-bias tradeoff can be quantified through the probabilistic bounds when we use $\Lambda^*_{\widetilde{\mathcal{S}}}(\theta(t), \theta(*))$ instead of $\Lambda_{\mathcal{S}}(\theta(t), \theta(*))$. In some cases, more data can be used to tighten or approximate the bound on $|\Lambda_{\mathcal{S}}(\theta(t), \theta(*)) - \Lambda^*_{\widetilde{\mathcal{S}}}(\theta(t), \theta(*))|$, so that inequality (5.32) can always be verified. Then the above procedure still yields an exact sampler. In many cases, however, so much data has to be used that the computation gains of subsampling are negligible. As such, we typically view this scheme as one with quantifiable bias. See [12] for further discussions and developments.

Overall, due to the similarities between our irreversible jump sampler and the MH algorithm, many methods developed specifically for MH can be applied in our context, which is quite appealing. For example, as discussed above, we have directly applied the subsampling approach designed for scaling MH to our approach. We can also combine our irreversible jump sampler with the RAM algorithm to further improve exploration in the case of multimodal targets.

5.8 Conclusion

In this chapter, we proposed frameworks for MCMC algorithms with both continuous dynamics and jump processes. We analyzed each of these components separately, and then showed how to combine them. For each component, we decomposed the dynamics into reversible and irreversible processes, and with a parameterization that was easy to specify while ensuring the correct stationary distribution.

First, we found that any continuous Markov process (with a mild integrability condition) can always be parameterized by its stationary distribution (i.e. the target distribution) $\pi(\mathbf{z})$, a positive semi-definite diffusion matrix $\mathbf{D}(\mathbf{z})$, and a skew-symmetric curl matrix $\mathbf{Q}(\mathbf{z})$. We analyzed the properties of the process in terms of $\mathbf{D}(\mathbf{z})$ and $\mathbf{Q}(\mathbf{z})$. In the context of Bayesian analysis with large datasets, we further discussed scalable methods using stochastic gradients.

Second, we turned to jump processes and considered a parameterization in terms of a symmetric kernel function $S(\mathbf{y}, \mathbf{z})$ and an anti-symmetric kernel function $A(\mathbf{y}, \mathbf{z})$. We showed that when $\int_{\mathbf{y}} A(\mathbf{y}, \mathbf{z}) d\mathbf{y} = 0$, the jump process has the target distribution $\pi(\mathbf{z})$ as its stationary distribution. When $A(\mathbf{y}, \mathbf{z}) \neq 0$, the jump process is irreversible. Facilitated by the framework, we constructed a new class of irreversible sampling algorithms that can be implemented similarly to the MH algorithm while directly satisfying $\int_{\mathbf{y}} A(\mathbf{y}, \mathbf{z}) d\mathbf{y} = 0$. Our experiments demonstrate that our proposed irreversible jump sampler is more efficient than the traditional reversible ones across a broad range of target distributions. We further discussed how a scalable variant is possible using the same subsampling idea as proposed for MH samplers.

Finally, we developed a technique to combine the continuous and jump processes by using the continuous dynamics as a proposal in the irreversible jump sampler. The directional effect of the continuous dynamics can facilitate better exploration of the target distribution than a simple proposal distribution. Likewise, one can also think of this framework as enabling a large step size to be taken in the continuous dynamic simulations while correcting for discretization error. We demonstrated that such a sampler can outperform samplers with independent proposals, samplers with continuous dynamics alone, or reversible versions of the combined approach (i.e., MALA).

The proposed framework requires a few critical choices to be made. For the continuous dynamics, we must specify $\mathbf{D}(\mathbf{z})$ and $\mathbf{Q}(\mathbf{z})$. For the jump processes, the specific algorithm we proposed requires selecting proposal distributions $\widetilde{f}(\mathbf{z}, \mathbf{z}^p | \mathbf{y}, \mathbf{y}^p)$ and $\widetilde{g}(\mathbf{z}, \mathbf{z}^p | \mathbf{y}, \mathbf{y}^p)$, and specifying the domain of the auxiliary variables $\mathbf{y}^p$. Our experiments have simply demonstrated that for certain choices of these matrices and parameters, we can achieve state-of-the-art performance in a variety of sampling tasks. An important direction for future work is to devise methods to analyze and explore the choices of these algorithmic parameters. For example, in higher dimensions, tuning the hyperparameters so that the irreversible jump sampler explores all dimensions efficiently is an interesting topic for further discussion and quantification. Also, in the irreversible MALA algorithm, we only used constant $\mathbf{D}(\mathbf{z})$ and $\mathbf{Q}(\mathbf{z})$, but using adaptive $\mathbf{D}(\mathbf{z})$ and $\mathbf{Q}(\mathbf{z})$ could potentially result in more efficient sampling algorithms [57, 102]. Another area of research is to examine using a higher order integration scheme in our irreversible MALA algorithm.

5.9 Future Directions

In this chapter, we proposed a reparametrization of the Markov processes that encoded the space of all correct samplers into matrices and transition kernels with straightforward conditions. Naturally, the next question is: for a given target distribution, is there an optimal choice of such parameters leading to the greatest mixing rate? This question is even harder than proposing a recipe that grants the correct stationary distribution, since the latter only concerns the eigenfunction corresponding to the greatest eigenvalue, the former asks about the gap between the first and the second eigenvalue. Calculating the spectral gap for a given Markov process can already be a very hard problem.

On another note, this chapter is mainly focused on the choice of Markov dynamics for MCMC algorithms. A very important step from Markov dynamics to practical sampling algorithms is the simulation of such dynamics. For example, different numerical schemes for the same continuous Markov dynamics can lead to vastly different performance in the resulting algorithm (from not even stable to accurate and efficient). This issue in the numerical analysis realm poses another area of exploration. In particular, traditional numerical analysis studies focus on constructing discretization schemes with more stability and accuracy. MCMC algorithms, on the other hand, is more concerned with the mixing rate than accuracy, since an accept-reject step can be applied. How to incorporate tools from numerical methods to MCMC algorithm is a novel and interesting question.

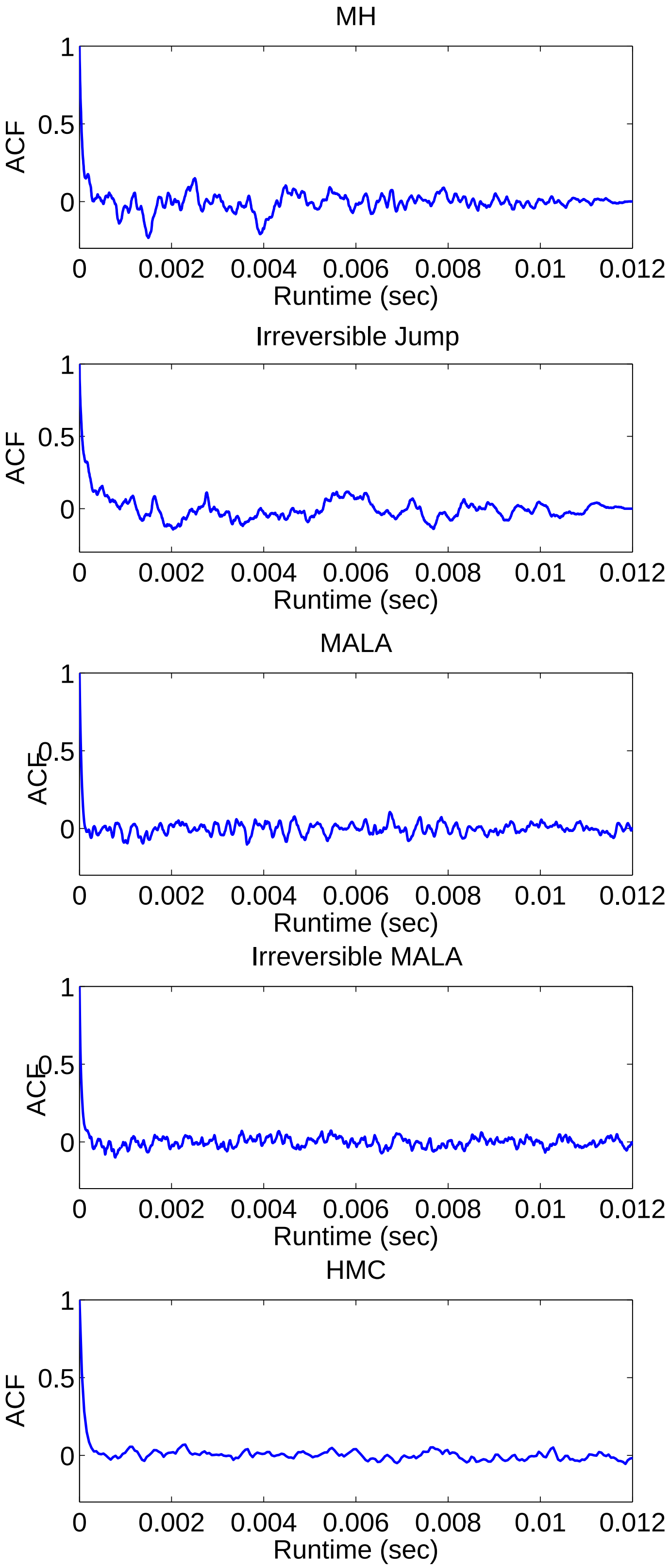


Figure 5.11: Comparison of the autocorrelation functions between different methods.

Chapter 6

SCALABLE BAYESIAN INFERENCE THROUGH STOCHASTIC GRADIENT MCMC

Many recent Markov chain Monte Carlo (MCMC) samplers leverage continuous dynamics to define a transition kernel that efficiently explores a target distribution. In tandem, a focus has been on devising scalable variants that subsample the data and use stochastic gradients in place of full-data gradients in the dynamic simulations. However, such stochastic gradient MCMC samplers have lagged behind their full-data counterparts in terms of the complexity of dynamics considered since proving convergence in the presence of the stochastic gradient noise is non-trivial. Even with simple dynamics, significant physical intuition is often required to modify the dynamical system to account for the stochastic gradient noise. In Sec. 6.1, we make use of results from Sec. 2.3 and 5.2 to provide a complete recipe for constructing stochastic gradient MCMC samplers based on continuous Markov processes specified via two matrices. Any continuous Markov process that provides samples from the target distribution can be written in our framework. We show how previous continuous-dynamic samplers can be trivially "reinvented" in our framework, avoiding the complicated sampler-specific proofs. We likewise use our recipe to straightforwardly propose a new state-adaptive sampler: *stochastic gradient Riemann Hamiltonian Monte Carlo* (SGRHMC). Our experiments on simulated data and a streaming Wikipedia analysis demonstrate that the proposed SGRHMC sampler inherits the benefits of Riemann HMC, with the scalability of stochastic gradient methods.

With stochastic gradient MCMC (SG-MCMC) algorithms successfully applied in Bayesian inference with large datasets with i.i.d data in Sec. 6.1, we further develop an SG-MCMC algorithm to learn the parameters of hidden Markov models (HMMs) for time-dependent data in Sec. 6.2. The challenge in applying SG-MCMC to dependent data is the need to break the dependencies when considering minibatches of observations. We propose an algorithm that harnesses the inherent memory decay of the process. We demonstrate the effectiveness of our algorithm on synthetic experiments and on an ion channel recording dataset. In terms of runtime, our algorithm significantly outperforms the corresponding batch MCMC algorithm.

6.1 A Complete Recipe for Stochastic Gradient MCMC

Recently, stochastic gradient variants of the continuous-dynamic-based samplers discussed in Sec. 5.2 have proven quite useful in scaling the methods to large datasets [182, 1, 28, 2, 37]. At each iteration, these samplers use data subsamples—or *minibatches*—rather than the full dataset. Stochastic gradient Langevin dynamics (SGLD) [182] innovated in this area by connecting stochastic optimization with a first-order Langevin dynamic MCMC technique, showing that adding the "right amount" of noise to stochastic gradient ascent iterates leads to samples from the target posterior as the step size is annealed. Stochastic gradient Hamiltonian Monte Carlo (SGHMC) [28] builds on this idea, but importantly incorporates the efficient exploration provided by the HMC momentum term. A key insight in that paper was that the naïve stochastic gradient variant of HMC actually leads to an incorrect stationary distribution (also see [14]); instead a modification to the dynamics underlying HMC is needed to account for the stochastic gradient noise. Variants of both SGLD and SGHMC with further modifications to improve efficiency have also recently been proposed [1, 121, 37].

In the plethora of past MCMC methods that explicitly leverage continuous dynamics—including HMC, Riemann manifold HMC, and the stochastic gradient methods—the focus has been on showing that the intricate dynamics leave the target posterior distribution invariant. Innovating in this arena requires constructing novel dynamics and simultaneously ensuring that the target distribution is the stationary distribution. This can be quite challenging, and often requires significant physical and geometrical intuition [28, 121, 37]. A natural question, then, is whether there exists a general recipe for devising such continuous-dynamic MCMC methods that naturally lead to invariance of the target distribution. In this section, we use results from Sec. 2.3 and 5.2 to answer this question to the affirmative. Furthermore, because any continuous Markov process admits a representation in the form of Eq. (5.6), we propose a *complete* recipe for stochastic gradient MCMC. That is, any valid stochastic gradient MCMC method can be cast within our framework, including SGLD, SGHMC, their recent variants, and any future developments in this area. Our method provides a unifying framework of past algorithms, as well as a practical tool for devising new samplers and testing the correctness of proposed samplers.

The same as in Sec. 5.2, the recipe involves defining a (stochastic) system parameterized by two matrices: a positive semidefinite diffusion matrix, $\mathbf{D}(\mathbf{z})$, and a skew-symmetric curl matrix, $\mathbf{Q}(\mathbf{z})$, where $\mathbf{z} = (\theta, r)$ with θ our model parameters of interest and r a set of auxiliary variables. The dynamics are then written explicitly in terms of the target stationary distribution and these two matrices. By varying the choices of $\mathbf{D}(\mathbf{z})$ and $\mathbf{Q}(\mathbf{z})$, we explore the space of MCMC methods that maintain the correct invariant distribution.

For any given $\mathbf{D}(\mathbf{z})$, $\mathbf{Q}(\mathbf{z})$, and target distribution, we provide practical algorithms for implementing either full-data or minibatch-based variants of the sampler. In Sec. 6.1.2, we cast many previous continuous-dynamic samplers in our framework, finding their $\mathbf{D}(\mathbf{z})$ and $\mathbf{Q}(\mathbf{z})$. We then show how these existing $\mathbf{D}(\mathbf{z})$ and $\mathbf{Q}(\mathbf{z})$ building blocks can be used to devise new samplers. In Sec. 6.1.2 we demonstrate our ability to construct new and relevant samplers by proposing *stochastic gradient Riemann Hamiltonian Monte Carlo*, the existence of which was previously only speculated. We demonstrate the utility of this sampler on synthetic data and in a streaming Wikipedia analysis using latent Dirichlet allocation [19].

6.1.1 Complete Stochastic Gradient MCMC Framework

We start with the standard MCMC goal of drawing samples from a target distribution, which we take to be the posterior $p(\theta|\mathcal{S})$ of model parameters $\theta \in \mathbb{R}^d$ given an observed dataset $\mathcal{S}$. Throughout, we assume i.i.d. data $\mathbf{x} \sim p(\mathbf{x}|\theta)$. We write $p(\theta|\mathcal{S}) \propto \exp(-U(\theta))$, with *potential function* $U(\theta) = -\sum_{\mathbf{x}\in\mathcal{S}} \log p(\mathbf{x}|\theta) - \log p(\theta)$. Algorithms like HMC [113, 57] further augment the space of interest with auxiliary variables r and sample from $p(\mathbf{z}|\mathcal{S}) \propto \exp(-H(\mathbf{z}))$, with *Hamiltonian*

$$H(\mathbf{z}) = H(\theta, r) = U(\theta) + g(\theta, r), \quad \text{such that} \int \exp(-g(\theta, r))\mathrm{d}r = constant. \qquad (6.1)$$

Marginalizing the auxiliary variables gives us the desired distribution on θ. In this section, we generically consider $\mathbf{z}$ as the samples we seek to draw; $\mathbf{z}$ could represent θ itself, or an augmented state space in which case we simply discard the auxiliary variables to perform the desired marginalization.

As in HMC, the idea is to translate the task of sampling from the posterior distribution to simulating from a continuous dynamical system which is used to define a Markov transition kernel. That is, over any interval h, the differential equation defines a mapping from the state at time t to the state at time $t+h$. One can then discuss the evolution of the distribution $p(\mathbf{z}, t)$ under the dynamics, as characterized by the Fokker-Planck equation for stochastic dynamics [147] or the Liouville equation for deterministic dynamics [191]. This evolution can be used to analyze the *invariant distribution* of the dynamics, $p^s(\mathbf{z})$. When considering deterministic dynamics, as in HMC, a jump process must be added to ensure ergodicity. If the resulting stationary distribution is equal to the target posterior, then simulating from the process can be equated with drawing samples from the posterior.

If the stationary distribution is *not* the target distribution, a Metropolis-Hastings (MH) correction can often be applied. Unfortunately, such correction steps require a costly computation on the entire dataset. Even if one can compute the MH correction, if the dynamics do not *nearly* lead to the correct stationary distribution, then the rejection rate can be high even for short simulation periods h. Furthermore, for many stochastic gradient MCMC samplers, computing the probability of the reverse path is infeasible, obviating the use of MH. As such, a focus in the literature is on defining dynamics with the right target distribution, especially in large-data scenarios where MH corrections are computationally burdensome or infeasible.

Devising SDEs with a Specified Target Stationary Distribution

Generically, all continuous Markov processes that one might consider for sampling can be written as a stochastic differential equation (SDE) of the form:

$$\mathrm{d}\mathbf{z} = \mathbf{f}(\mathbf{z})\mathrm{d}t + \sqrt{2\mathbf{D}(\mathbf{z})}\mathrm{d}\mathbf{W}(t), \tag{6.2}$$

where $\mathbf{f}(\mathbf{z})$ denotes the deterministic drift and often relates to the gradient of $H(\mathbf{z})$, $\mathbf{W}(t)$ is a d-dimensional Wiener process, and $\mathbf{D}(\mathbf{z})$ is a positive semidefinite diffusion matrix. Clearly, however, not all choices of $\mathbf{f}(\mathbf{z})$ and $\mathbf{D}(\mathbf{z})$ yield the stationary distribution $p^s(\mathbf{z}) \propto \exp(-H(\mathbf{z}))$.

When $\mathbf{D}(\mathbf{z}) = 0$, as in HMC, the dynamics of Eq. (6.2) become deterministic. Our exposition focuses on SDEs, but our analysis applies to deterministic dynamics as well. In this case, our framework—using the Liouville equation in place of Fokker-Planck—ensures that the deterministic dynamics leave the target distribution invariant. For ergodicity, a jump process must be added, which is not considered in our recipe, but tends to be straightforward (e.g., momentum resampling in HMC).

To devise a recipe for constructing SDEs with the correct stationary distribution, we make use of the results in Sec. 2.3 and 5.2 and propose writing $\mathbf{f}(\mathbf{z})$ directly in terms of the target distribution:

$$\mathbf{f}(\mathbf{z}) = -\big[\mathbf{D}(\mathbf{z}) + \mathbf{Q}(\mathbf{z})\big]\nabla H(\mathbf{z}) + \Gamma(\mathbf{z}), \quad \Gamma_i(\mathbf{z}) = \sum_{j=1}^{d} \frac{\partial}{\partial \mathbf{z}_j}\big(\mathbf{D}_{ij}(\mathbf{z}) + \mathbf{Q}_{ij}(\mathbf{z})\big). \tag{6.3}$$

Here, $\mathbf{Q}(\mathbf{z})$ is a skew-symmetric curl matrix representing the deterministic traversing effects seen in HMC procedures. In contrast, the diffusion matrix $\mathbf{D}(\mathbf{z})$ determines the strength of the Wiener-process-driven diffusion. Matrices $\mathbf{D}(\mathbf{z})$ and $\mathbf{Q}(\mathbf{z})$ can be adjusted to attain faster convergence to the posterior distribution.

Importantly, as discussed in Sec. 2.3 and 5.2, sampling the stochastic dynamics of Eq. (6.2) (according to Itô integral) with $\mathbf{f}(\mathbf{z})$ as in Eq. (6.3) leads to the desired posterior distribution as the stationary distribution: $p^s(\mathbf{z}) \propto \exp(-H(\mathbf{z}))$, for any choice of positive semidefinite $\mathbf{D}(\mathbf{z})$ and skew-symmetric $\mathbf{Q}(\mathbf{z})$ assuming the process is ergodic. Also, for any continuous Markov process with the desired stationary distribution, $p^s(\mathbf{z})$, there exists an SDE as in Eq. (6.2) with $\mathbf{f}(\mathbf{z})$ as in Eq. (6.3).

A Practical Algorithm

In practice, simulation relies on an ϵ-discretization of the SDE, leading to a **full-data** update rule

$$\mathbf{z}_{t+1} \leftarrow \mathbf{z}_t - \epsilon_t \left[\big(\mathbf{D}(\mathbf{z}_t) + \mathbf{Q}(\mathbf{z}_t)\big)\nabla H(\mathbf{z}_t) + \Gamma(\mathbf{z}_t)\right] + \mathcal{N}(0, 2\epsilon_t \mathbf{D}(\mathbf{z}_t)). \tag{6.4}$$

Calculating the gradient of $H(\mathbf{z})$ involves evaluating the gradient of $U(\theta)$. For a stochastic gradient method, the assumption is that $U(\theta)$ is too computationally intensive to compute as it relies on a sum over all data points (see Sec. 6.1.1). Instead, such stochastic gradient algorithms examine *independently sampled* data subsets $\widetilde{\mathcal{S}} \subset \mathcal{S}$ and the corresponding potential for these data:

$$\widetilde{U}(\theta) = -\frac{|\mathcal{S}|}{|\widetilde{\mathcal{S}}|} \sum_{\mathbf{x} \in \widetilde{\mathcal{S}}} \log p(\mathbf{x}|\theta) - \log p(\theta); \quad \widetilde{\mathcal{S}} \subset \mathcal{S}. \tag{6.5}$$

The specific form of Eq. (6.22) implies that $\widetilde{U}(\theta)$ is an unbiased estimator of $U(\theta)$. As such, a gradient computed based on $\widetilde{U}(\theta)$—called a *stochastic gradient* [148]—is a noisy, but unbiased estimator of the full-data gradient. The key question in many of the existing stochastic gradient MCMC algorithms is whether the noise injected by the stochastic gradient adversely affects the stationary distribution of the modified dynamics (using $\nabla\widetilde{U}(\theta)$ in place of $\nabla U(\theta)$). One way to analyze the impact of the stochastic gradient is to make use of the central limit theorem and assume

$$\nabla\widetilde{U}(\theta) = \nabla U(\theta) + \mathcal{N}(0, \mathbf{V}(\theta)), \tag{6.6}$$

resulting in a noisy Hamiltonian gradient $\nabla\widetilde{H}(\mathbf{z}) = \nabla H(\mathbf{z}) + [\mathcal{N}(0, \mathbf{V}(\theta)), \mathbf{0}]^T$. Simply plugging in $\nabla\widetilde{H}(\mathbf{z})$ in place of $\nabla H(\mathbf{z})$ in Eq. (6.4) results in dynamics with an additional noise term $\big(\mathbf{D}(\mathbf{z}_t) + \mathbf{Q}(\mathbf{z}_t)\big)[\mathcal{N}(0, \mathbf{V}(\theta)), \mathbf{0}]^T$. To counteract this, assume we have an estimate $\hat{\mathbf{B}}_t$ of the variance of this additional noise satisfying $2\mathbf{D}(\mathbf{z}_t) - \epsilon_t\hat{\mathbf{B}}_t \succeq 0$ (i.e., positive semidefinite). With small ϵ, this is always true since the stochastic gradient noise scales down faster than the added noise. Then, we can attempt to account for the stochastic gradient noise by simulating

$$\mathbf{z}_{t+1} \leftarrow \mathbf{z}_t - \epsilon_t \left[\big(\mathbf{D}(\mathbf{z}_t) + \mathbf{Q}(\mathbf{z}_t) \big) \nabla \widetilde{H}(\mathbf{z}_t) + \Gamma(\mathbf{z}_t) \right] + \mathcal{N}(0, \epsilon_t (2\mathbf{D}(\mathbf{z}_t) - \epsilon_t \hat{\mathbf{B}}_t)). \quad (6.7)$$

This provides our **stochastic gradient**—or ***minibatch***— variant of the sampler. In Eq. (6.24), the noise introduced by the stochastic gradient is multiplied by ϵ_t (and the compensation by ϵ_t^2), implying that the discrepancy between these dynamics and those of Eq. (6.4) approaches zero as ϵ_t goes to zero. As such, in this infinitesimal step size limit, since Eq. (6.4) yields the correct invariant distribution, so does Eq. (6.24). This avoids the need for a costly or potentially intractable MH correction. However, having to decrease ϵ_t to zero comes at the cost of increasingly small updates. We can also use a finite, small step size in practice, resulting in a biased (but faster) sampler. A similar bias-speed tradeoff was used in [82, 11] to construct MH samplers, in addition to being used in SGLD and SGHMC.

6.1.2 Applying the Theory to Construct Samplers

Casting Previous MCMC Algorithms within the Proposed Framework

We explicitly state how some recently developed MCMC methods fall within the proposed framework based on specific choices of $\mathbf{D}(\mathbf{z})$, $\mathbf{Q}(\mathbf{z})$ and $H(\mathbf{z})$ in Eq. (6.2) and (6.3). For the stochastic gradient methods, we show how our framework can be used to "reinvent" the samplers by guiding their construction and avoiding potential mistakes or inefficiencies caused by naïve implementations.

Hamiltonian Monte Carlo (HMC) The key ingredient in HMC [38, 113] is Hamiltonian dynamics, which simulate the physical motion of an object with position θ, momentum r, and mass $\mathbf{M}$ on an frictionless surface as follows (typically, a leapfrog step is used instead of the Euler-Maruyama integral shown in Eq. (6.4)):

$$\begin{cases} r_{t+1/2} \leftarrow r_t - \epsilon_t \nabla U(\theta_t)/2 \theta_{t+1} \leftarrow \theta_t + \epsilon_t \mathbf{M}^{-1} r_t \\ r_{t+1} \leftarrow r_{t+1/2} - \epsilon_t \nabla U(\theta_t)/2. \end{cases} \quad (6.8)$$

Eq. (6.8) is a special case of the proposed framework with $\mathbf{z} = (\theta, r)$, $H(\theta, r) = U(\theta) + \frac{1}{2} r^T M^{-1} r$, $\mathbf{Q}(\theta, r) = \begin{pmatrix} 0 & -\mathbf{I} \\ \mathbf{I} & 0 \end{pmatrix}$ and $\mathbf{D}(\theta, r) = \mathbf{0}$.

Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) As discussed in [28], simply replacing $\nabla U(\theta)$ by the stochastic gradient $\nabla\widetilde{U}(\theta)$ in Eq. (6.8) results in the following updates:

$$\text{Naive}: \begin{cases} \theta_{t+1} \leftarrow \theta_t + \epsilon_t \mathbf{M}^{-1} r_t \\ r_{t+1} \leftarrow r_t - \epsilon_t \nabla\widetilde{U}(\theta_t) \approx r_t - \epsilon_t \nabla U(\theta_t) + \mathcal{N}(0, \epsilon_t^2 \mathbf{V}(\theta_t)), \end{cases} \tag{6.9}$$

where the $\approx$ arises from the approximation of Eq. (6.23). Careful study shows that Eq. (6.9) *cannot* be rewritten into our proposed framework, which hints that such a naïve stochastic gradient version of HMC is not correct. Interestingly, the authors of [28] proved that this naïve version indeed does not have the correct stationary distribution. In our framework, we see that the noise term $\mathcal{N}(0, 2\epsilon_t \mathbf{D}(\mathbf{z}))$ is paired with a $\mathbf{D}(\mathbf{z})\nabla H(\mathbf{z})$ term, hinting that such a term should be added to Eq. (6.9). Here, $\mathbf{D}(\theta, r) = \begin{pmatrix} 0 & 0 \\ 0 & \epsilon \mathbf{V}(\theta) \end{pmatrix}$, which means we need to add $\mathbf{D}(\mathbf{z})\nabla H(\mathbf{z}) = \epsilon \mathbf{V}(\theta) \nabla_r H(\theta, r) = \epsilon \mathbf{V}(\theta) \mathbf{M}^{-1} r$. Interestingly, this is the correction strategy proposed in [28], but through a physical interpretation of the dynamics. In particular, the term $\epsilon \mathbf{V}(\theta)\mathbf{M}^{-1} r$ (or, generically, $\mathbf{C}\mathbf{M}^{-1} r$ where $\mathbf{C} \succeq \epsilon \mathbf{V}(\theta)$) has an interpretation as friction and leads to second order Langevin dynamics:

$$\begin{cases} \theta_{t+1} \leftarrow \theta_t + \epsilon_t \mathbf{M}^{-1} r_t \\ r_{t+1} \leftarrow r_t - \epsilon_t \nabla\widetilde{U}(\theta_t) - \epsilon_t \mathbf{C}\mathbf{M}^{-1} r_t + \mathcal{N}(0, \epsilon_t(2\mathbf{C} - \epsilon_t \hat{\mathbf{B}}_t)). \end{cases} \tag{6.10}$$

Here, $\hat{\mathbf{B}}_t$ is an estimate of $\mathbf{V}(\theta_t)$. This method now fits into our framework with $H(\theta, r)$ and $\mathbf{Q}(\theta, r)$ as in HMC, but with $\mathbf{D}(\theta, r) = \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{C} \end{pmatrix}$. This example shows how our theory can be used to identify invalid samplers and provide guidance on how to effortlessly correct the mistakes; this is crucial when physical intuition is not available. Once the proposed sampler is cast in our framework with a specific $\mathbf{D}(\mathbf{z})$ and $\mathbf{Q}(\mathbf{z})$, there is no need for sampler-specific proofs, such as those of [28].

Stochastic Gradient Langevin Dynamics (SGLD) SGLD [182] proposes to use the following first order (no momentum) Langevin dynamics to generate samples

$$\theta_{t+1} \leftarrow \theta_t - \epsilon_t \mathbf{D} \nabla\widetilde{U}(\theta_t) + \mathcal{N}(0, 2\epsilon_t \mathbf{D}). \tag{6.11}$$

This algorithm corresponds to taking $\mathbf{z} = \theta$ with $H(\theta) = U(\theta)$, $\mathbf{D}(\theta) = \mathbf{D}$, $\mathbf{Q}(\theta) = \mathbf{0}$, and $\hat{\mathbf{B}}_t = \mathbf{0}$. As motivated by Eq. (6.24) of our framework, the variance of the stochastic gradient can be subtracted from the sampler injected noise to make the finite stepsize simulation more accurate. This variant of SGLD leads to the stochastic gradient Fisher scoring algorithm [1].

Stochastic Gradient Riemannian Langevin Dynamics (SGRLD) SGLD can be generalized to use an adaptive diffusion matrix $\mathbf{D}(\theta)$. Specifically, it is interesting to take $\mathbf{D}(\theta) = \mathbf{G}^{-1}(\theta)$, where $\mathbf{G}(\theta)$ is the Fisher information metric. The sampler dynamics are given by

$$\theta_{t+1} \leftarrow \theta_t - \epsilon_t[\mathbf{G}(\theta_t)^{-1}\nabla\widetilde{U}(\theta_t) + \Gamma(\theta_t)] + \mathcal{N}(0, 2\epsilon_t\mathbf{G}(\theta_t)^{-1}). \tag{6.12}$$

Taking $\mathbf{D}(\theta) = \mathbf{G}(\theta)^{-1}$, $\mathbf{Q}(\theta) = \mathbf{0}$, and $\hat{\mathbf{B}}_t = \mathbf{0}$, this SGRLD [121] method falls into our framework with correction term $\Gamma_i(\theta) = \sum_j \frac{\partial \mathbf{D}_{ij}(\theta)}{\partial \theta_j}$. It is interesting to note that in earlier literature [**?**], $\Gamma_i(\theta)$ was taken to be $2\,|\mathbf{G}(\theta)|^{-1/2}\sum_j \frac{\partial}{\partial \theta_j}\left(\mathbf{G}_{ij}^{-1}(\theta)|\mathbf{G}(\theta)|^{1/2}\right)$. More recently, it was found that this correction term corresponds to the distribution function with respect to a non-Lebesgue measure [150]; for the Lebesgue measure, the revised $\Gamma_i(\theta)$ was as determined by our framework [150]. Again, we have an example of our theory providing guidance in devising correct samplers.

Stochastic Gradient Nosé-Hoover Thermostat (SGNHT) Finally, the SGNHT [37] method incorporates ideas from thermodynamics to further increase adaptivity by augmenting the SGHMC system with an additional scalar auxiliary variable, ξ. The algorithm uses the following dynamics:

$$\begin{cases} \theta_{t+1} \leftarrow \theta_t + \epsilon_t r_t \\ r_{t+1} \leftarrow r_t - \epsilon_t \nabla\widetilde{U}(\theta_t) - \epsilon_t\xi_t r_t + \mathcal{N}(0, \epsilon_t(2A - \epsilon_t\hat{\mathbf{B}}_t)) \\ \xi_{t+1} \leftarrow \xi_t + \epsilon_t\left(\frac{1}{d}r_t^T r_t - 1\right). \end{cases} \tag{6.13}$$

We can take $\mathbf{z} = (\theta, r, \xi)$, $H(\theta, r, \xi) = U(\theta) + \frac{1}{2}r^T r + \frac{1}{2d}(\xi - A)^2$, $\mathbf{D}(\theta, r, \xi) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & A\cdot\mathbf{I} & 0 \\ 0 & 0 & 0 \end{pmatrix}$, and $\mathbf{Q}(\theta, r, \xi) = \begin{pmatrix} 0 & -\mathbf{I} & 0 \\ \mathbf{I} & 0 & r/d \\ 0 & -r^T/d & 0 \end{pmatrix}$ to place these dynamics within our framework.

Summary In our framework, SGLD and SGRLD take $\mathbf{Q}(\mathbf{z}) = 0$ and instead stress the design of the diffusion matrix $\mathbf{D}(\mathbf{z})$, with SGLD using a constant $\mathbf{D}(\mathbf{z})$ and SGRLD an adaptive, θ-dependent diffusion matrix to better account for the geometry of the space being explored. On the other hand, HMC takes $\mathbf{D}(\mathbf{z}) = 0$ and focuses on the curl matrix $\mathbf{Q}(\mathbf{z})$. SGHMC combines SGLD with HMC through non-zero $\mathbf{D}(\theta)$ and $\mathbf{Q}(\theta)$ matrices. SGNHT then extends SGHMC by taking $\mathbf{Q}(\mathbf{z})$ to be state dependent. The relationships between these methods are depicted in the Supplement, which likewise contains a discussion of the tradeoffs between these two matrices. In short, $\mathbf{D}(\mathbf{z})$ can guide escaping from local modes while $\mathbf{Q}(\mathbf{z})$ can enable rapid traversing of low-probability regions, especially when state adaptation is incorporated. We readily see that most of the product space $\mathbf{D}(\mathbf{z}) \times \mathbf{Q}(\mathbf{z})$, defining the space of all possible samplers, has yet to be filled.

A lot of choices of $\mathbf{D}(\mathbf{z})$ and $\mathbf{Q}(\mathbf{z})$ could potentially result in faster convergence of the samplers than those previously explored. For example, $\mathbf{D}(\mathbf{z})$ determines how much noise is introduced. Hence, an adaptive diffusion matrix $\mathbf{D}(\mathbf{z})$ can facilitate a faster escape from a local mode if $||\mathbf{D}(\mathbf{z})||$ is larger in regions of low probability, and can increase accuracy near the global mode if $||\mathbf{D}(\mathbf{z})||$ is smaller in regions of high probability. Motivated by the fact that a majority of the parameter space is covered by low probability mass regions where less accuracy is often needed, one might want to traverse these regions quickly. As such, an adaptive curl matrix $\mathbf{Q}(\mathbf{z})$ with 2-norm growing with the level set of the distribution can facilitate a more efficient sampler. We explore an example of this in the gSGRHMC algorithm of the synthetic experiments (see Sec. 6.1.3).

Stochastic Gradient Riemann Hamiltonian Monte Carlo

In Sec. 6.1.2, we have shown how our framework unifies existing samplers. In this section, we now use our framework to guide the development of a new sampler. While SGHMC [28] inherits the momentum term of HMC, making it easier to traverse the space of parameters, the underlying geometry of the target distribution is still not utilized. Such information can usually be represented by the Fisher information metric [**?**] or local Hessian information, denoted as $\mathbf{G}(\theta)$, which can be used to precondition the dynamics. For our proposed system, we consider $H(\theta, r) = U(\theta) + \frac{1}{2}r^T r$, as in HMC/SGHMC methods, and modify the $\mathbf{D}(\theta, r)$ and $\mathbf{Q}(\theta, r)$ of SGHMC to account for the geometry as follows:

$$\mathbf{D}(\theta, r) = \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{G}(\theta)^{-1} \end{pmatrix}; \qquad \mathbf{Q}(\theta, r) = \begin{pmatrix} 0 & -\mathbf{G}(\theta)^{-1/2} \\ \mathbf{G}(\theta)^{-1/2} & 0 \end{pmatrix}.$$

We refer to this algorithm as *stochastic gradient Riemann Hamiltonian Monte Carlo* (**SGRHMC**). Our theory holds for any positive definite $\mathbf{G}(\theta)$, yielding a *generalized SGRHMC* (**gSGRHMC**) algorithm, which can be helpful when the Fisher information metric is hard to compute.

Algorithm 6: Generalized Stochastic Gradient Riemann Hamiltonian Monte Carlo

```
initialize $(\theta_0, r_0)$
for $t = 0, 1, 2 \cdots$ do
    optionally, periodically resample momentum $r$ as $r^{(t)} \sim \mathcal{N}(0, \mathbf{I})$
    $\theta_{t+1} \leftarrow \theta_t + \epsilon_t \mathbf{G}(\theta_t)^{-1/2} r_t, \quad \Sigma_t \leftarrow \epsilon_t (2\mathbf{G}(\theta_t)^{-1} - \epsilon_t \hat{\mathbf{B}}_t)$
    $r_{t+1} \leftarrow r_t - \epsilon_t \mathbf{G}(\theta_t)^{-1/2} \nabla_\theta \widetilde{U}(\theta_t) + \epsilon_t \nabla_\theta(\mathbf{G}(\theta_t)^{-1/2}) - \epsilon_t \mathbf{G}(\theta_t)^{-1} r_t + \mathcal{N}\Big(0, \Sigma_t\Big)$
end
```

A naïve implementation of a state-dependent SGHMC algorithm might simply (i) precondition the HMC update, (ii) replace $\nabla U(\theta)$ by $\nabla \widetilde{U}(\theta)$, and (iii) add a state-dependent friction term on the order of the diffusion matrix to counterbalance the noise as in SGHMC, resulting in:

$$\text{Naive}: \begin{cases} \theta_{t+1} \leftarrow & \theta_t + \epsilon_t \mathbf{G}(\theta_t)^{-1/2} r_t \\ r_{t+1} \leftarrow & r_t - \epsilon_t \mathbf{G}(\theta_t)^{-1/2} \nabla_\theta \widetilde{U}(\theta_t) - \epsilon_t \mathbf{G}(\theta_t)^{-1} r_t + \mathcal{N}(0, \epsilon_t (2\mathbf{G}(\theta_t)^{-1} - \epsilon_t \hat{\mathbf{B}}_t)). \end{cases} \tag{6.14}$$

However, as we show in Sec. 6.1.3, samples from these dynamics do not converge to the desired distribution. Indeed, this system cannot be written within our framework. Instead, we can simply follow our framework and, as indicated by Eq. (6.24), consider the following update rule:

$$\begin{cases} \theta_{t+1} \leftarrow \theta_t + \epsilon_t \mathbf{G}(\theta_t)^{-1/2} r_t \\ r_{t+1} \leftarrow r_t - \epsilon_t [\mathbf{G}(\theta)^{-1/2} \nabla_\theta \widetilde{U}(\theta_t) + \nabla_\theta \left(\mathbf{G}(\theta_t)^{-1/2}\right) - \mathbf{G}(\theta_t)^{-1} r_t] + \mathcal{N}(0, \epsilon_t (2\mathbf{G}(\theta_t)^{-1} - \epsilon_t \hat{\mathbf{B}}_t)), \end{cases} \tag{6.15}$$

which includes a correction term $\nabla_\theta \left(\mathbf{G}(\theta)^{-1/2}\right)$, with i-th component $\sum_j \frac{\partial}{\partial \theta_j} \left(\mathbf{G}(\theta)^{-1/2}\right)_{ij}$. The practical implementation of gSGRHMC is outlined in Algorithm 6.

6.1.3 Experiments

In Sec. 6.1.3, we show that gSGRHMC can excel at rapidly exploring distributions with complex landscapes. We then apply SGRHMC to sampling in a latent Dirichlet allocation (LDA) model on a large Wikipedia dataset in Sec. 6.1.3.

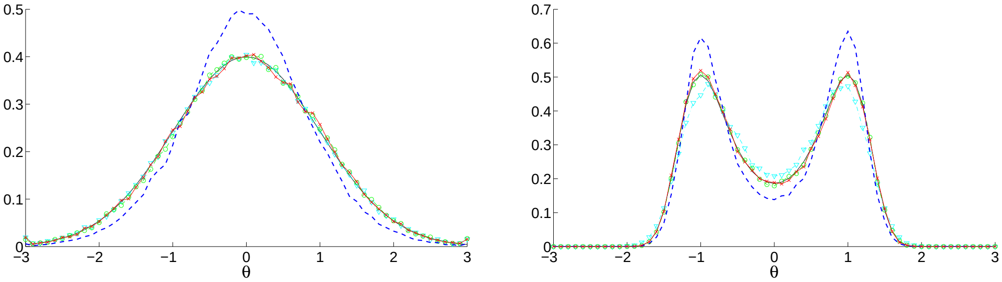


Figure 6.1: For two simulated 1D distributions (black) defined by $U(\theta) = \theta^2/2$ (*left*) and $U(\theta) = \theta^4 - 2\theta^2$ (*right*), comparison of SGLD, SGHMC, the naïve SGRHMC of Eq. (6.14), and the gSGRHMC of Eq. (6.15).

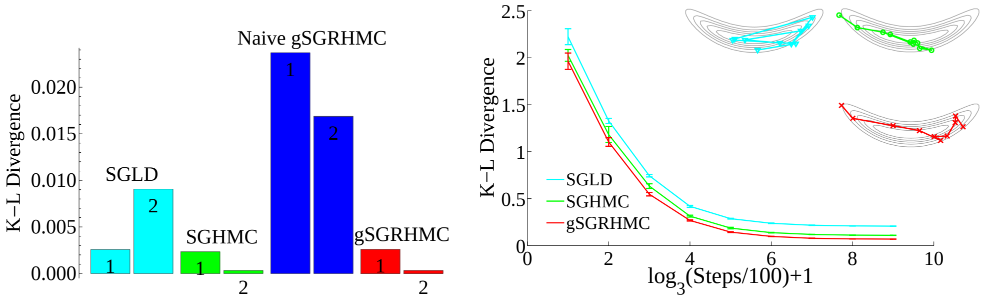


Figure 6.2: *Left:* For two simulated 1D distributions defined by $U(\theta) = \theta^2/2$ (*one peak*) and $U(\theta) = \theta^4 - 2\theta^2$ (*two peaks*), we compare the KL divergence of methods: SGLD, SGHMC, the naïve SGRHMC of Eq. (6.14), and the gSGRHMC of Eq. (6.15) relative to the true distribution in each scenario (left and right bars labeled by 1 and 2). *Right:* For a correlated 2D distribution with $U(\theta_1, \theta_2) = \theta_1^4/10 + (4\cdot(\theta_2+1.2) - \theta_1^2)^2/2$, we see that our gSGRHMC most rapidly explores the space relative to SGHMC and SGLD. Contour plots of the distribution along with paths of the first 10 sampled points are shown for each method.

Synthetic Experiments

In this section we aim to empirically (i) validate the correctness of our recipe and (ii) assess the effectiveness of gSGRHMC. In Fig. 6.5(left), we consider two univariate distributions (shown in the Supplement) and compare SGLD, SGHMC, the naïve state-adaptive SGHMC of Eq. (6.14), and our proposed gSGRHMC of Eq. (6.15). We specifically consider

	Original LDA	**Expanded Mean**
Parameter θ	$\beta_{kw} = \theta_{kw}$	$\beta_{kw} = \frac{\theta_{kw}}{\sum_w \theta_{kw}}$
Prior $p(\theta)$	$p(\theta_k) = \mathrm{Dir}(\alpha)$	$p(\theta_{kw}) = \Gamma(\alpha, 1)$

Method	**Average Runtime per 100 Docs**
SGLD	0.778s
SGHMC	0.815s
SGRLD	0.730s
SGRHMC	0.806s

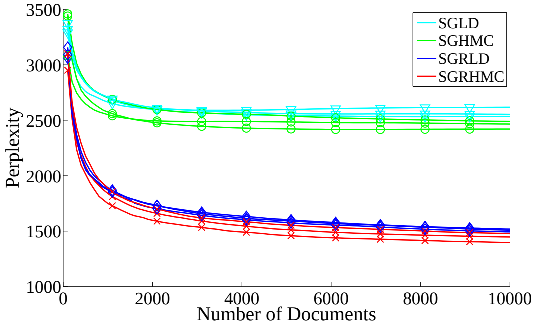


Figure 6.3: *Upper Left:* Expanded mean parameterization of the LDA model. *Lower Left:* Average runtime per 100 Wikipedia entries for all methods. *Right:* Perplexity versus number of Wikipedia entries processed.

$\mathbf{G}(\theta)^{-1} = D\sqrt{|\widetilde{U}(\theta) + C|}$. The constant C ensures that $\widetilde{U}(\theta) + C$ is positive in most cases so that the fluctuation is indeed smaller when the probability density function is higher. Note that we define $\mathbf{G}(\theta)$ in terms of $\widetilde{U}(\theta)$ to avoid a costly full-data computation. We choose $D = 1.5$ and $C = 0.5$ in the experiments. The design of $\mathbf{G}$ is motivated by the discussion in Sec. 6.1.2, taking $\mathbf{Q}(\theta)$ to have 2-norm growing with the level sets of the potential function can lead to faster exploration of the posterior.

As expected, the naïve implementation does not converge to the target distribution. In contrast, the gSGRHMC algorithm obtained via our recipe indeed has the correct invariant distribution and efficiently explores the distributions. In the second experiment, we sample a bivariate distribution with strong correlation. The results are shown in Fig. 6.5(right). The comparison between SGLD, SGHMC, and our gSGRHMC method shows that both a state-dependent preconditioner and Hamiltonian dynamics help to make the sampler more efficient than either element on its own.

Online Latent Dirichlet Allocation

We also applied SGRHMC (with $\mathbf{G}(\theta) = \mathrm{diag}(\theta)^{-1}$, the Fisher information metric) to an *online* latent Dirichlet allocation (LDA) [19] analysis of topics present in Wikipedia entries. In LDA, each topic is associated with a distribution over words, with β_{kw} the probability of word w under topic k. Each document is comprised of a mixture of topics, with $\pi_k^{(d)}$

the probability of topic k in document d. Documents are generated by first selecting a topic $z_j^{(d)} \sim \pi^{(d)}$ for the jth word and then drawing the specific word from the topic as $x_j^{(d)} \sim \beta_{z_j^{(d)}}$. Typically, $\pi^{(d)}$ and β_k are given Dirichlet priors.

The goal of our analysis here is inference of the corpus-wide topic distributions β_k. Since the Wikipedia dataset is large and continually growing with new articles, it is not practical to carry out this task over the whole dataset. Instead, we scrape the corpus from Wikipedia in a streaming manner and sample parameters based on minibatches of data. Following the approach in [121], we first analytically marginalize the document distributions $\pi^{(d)}$ and, to resolve the boundary issue posed by the Dirichlet posterior of β_k defined on the probability simplex, use an expanded mean parameterization shown in Figure 6.3(upper left). Under this parameterization, we then compute $\nabla \log p(\theta|\mathbf{x})$ and, in our implementation, use boundary reflection to ensure the positivity of parameters θ_{kw}. The necessary expectation over word-specific topic indicators $z_j^{(d)}$ is approximated using Gibbs sampling separately on each document, as in [121].

We used minibatches of 50 documents and $K = 50$ topics. Similar to [121], the stochastic gradient of the log posterior of the parameter θ on a minibatch $\widetilde{\mathcal{S}}$ is calculated as

$$\frac{\partial \log p(\theta|\mathbf{x}, \alpha, \gamma)}{\partial \theta_{kw}} \approx \frac{\alpha - 1}{\theta_{kw}} - 1 + \frac{|\mathcal{S}|}{|\widetilde{\mathcal{S}}|} \sum_{d \in \widetilde{\mathcal{S}}} \mathbb{E}_{\mathbf{z}^{(d)}|\mathbf{x}^{(d)},\theta,\gamma} \left[\frac{n_{dkw}}{\theta_{kw}} - \frac{n_{dk\cdot}}{\theta_{k\cdot}} \right], \tag{6.16}$$

where α is the hyper-parameter for the Gamma prior of per-topic word distributions, and γ for the per-document topic distributions. Here, n_{dkw} is the count of how many times word w is assigned to topic k in document d (via $z_j^{(d)} = k$ for $x_j = w$). The $\cdot$ notation indicates $n_{dk\cdot} = \sum_w n_{dkw}$. To calculate the expectation of the latent topic assignment counts n_{dkw}, Gibbs sampling is used on the topic assignments in each document separately, using the conditional distributions

$$p(z_j^{(d)} = k|\mathbf{x}^{(d)}, \theta, \gamma) = \frac{\left(\gamma + n_{dk\cdot}^{\backslash j}\right) \theta_{kx_j^{(d)}}}{\sum_k \left(\gamma + n_{dk\cdot}^{\backslash j}\right) \theta_{kx_j^{(d)}}}, \tag{6.17}$$

where $\backslash j$ represents a count excluding the topic assignment variable $z_j^{(d)}$ being updated. See [121] for further details.

We follow the experimental settings in [121] for Riemmanian samplers (SGRLD and SGRHMC), taking the hyper-parameters of Dirichlet priors to be $\gamma = 0.01$ and $\alpha = 0.0001$. Since the non-Riemmanian samplers (SGLD and SGHMC) do not handle distributions with mass concentrated over small regions as well as the Riemmanian samplers, we found $\gamma = 0.1$ and $\alpha = 0.01$ to be optimal hyper-parameters for them and use these instead for SGLD and SGHMC. In doing so, we are modifying the posterior being sampled, but wished to provide as good of performance as possible for these baseline methods for a fair comparison. For the SGRLD method, we keep the stepsize schedule of $\epsilon_t = \left(a \cdot \left(1 + \frac{t}{b}\right)\right)^{-c}$ and corresponding optimal parameters a, b, c used in the experiment of [121]. For the other methods, we use a constant stepsize because it was easier to tune. (A constant stepsize for SGRLD performed worse than the schedule described above, so again we are trying to be as fair to baseline methods as possible when using non-constant stepsize for SGRLD.) A grid search is performed to find $\epsilon_t = 0.02$ for the SGRHMC method; $\epsilon_t = 0.01$, $\mathbf{D} = I$ (corresponding to Eq. (6.11)) for the SGLD method; and $\epsilon_t = 0.1$, $\mathbf{C} = \mathbf{M} = I$ (corresponding to Eq. (6.10)) for the SGHMC method.

For a randomly selected subset of topics, in Table 6.1 we show the top seven most heavily weighted words in the topic learned with the SGRHMC sampler.

“ENGINES”	speed	product	introduced	designs	fuel	quality
“ROYAL”	britain	queen	sir	earl	died	house
“ARMY”	commander	forces	war	general	military	colonel
“STUDY”	analysis	space	program	user	research	developed
“PARTY”	act	office	judge	justice	legal	vote
“DESIGN”	size	glass	device	memory	engine	cost
“PUBLIC”	report	health	community	industry	conference	congress
“CHURCH”	prayers	communion	religious	faith	historical	doctrine
“COMPANY”	design	production	produced	management	market	primary
“PRESIDENT”	national	minister	trial	states	policy	council
“SCORE”	goals	team	club	league	clubs	years

Table 6.1: The top seven most heavily weighted words (columns) associated with each of a randomly selected set of 11 topics (rows) learned with the SGRHMC sampler from 10,000 documents (about 0.3% of the articles in Wikipedia). The capitalized words in the first column represent the most heavily weighted word in each topic, and are used as the topic labels.

For all the methods, we report results of three random runs. When sampling distributions with mass concentrated over small regions, as in this application, it is important to incorporate geometric information via a Riemannian sampler [121]. The results in

Fig. 6.3(right) indeed demonstrate the importance of Riemannian variants of the stochastic gradient samplers. However, there also appears to be some benefits gained from the incorporation of the HMC term for both the Riemmannian and non-Reimannian samplers. The average runtime for the different methods are similar (see Fig. 6.3(lower left)) since the main computational bottleneck is the gradient evaluation. Overall, this application serves as an important example of where our newly proposed sampler can have impact.

6.1.4 Conclusion

We presented a general recipe for devising MCMC samplers based on continuous Markov processes. Our framework constructs an SDE specified by two matrices, a positive semidefinite $\mathbf{D}(\mathbf{z})$ and a skew-symmetric $\mathbf{Q}(\mathbf{z})$. We prove that for any $\mathbf{D}(\mathbf{z})$ and $\mathbf{Q}(\mathbf{z})$, we can devise a continuous Markov process with a specified stationary distribution. We also prove that for any continuous Markov process with the target stationary distribution, there exists a $\mathbf{D}(\mathbf{z})$ and $\mathbf{Q}(\mathbf{z})$ that cast the process in our framework. Our recipe is particularly useful in the more challenging case of devising stochastic gradient MCMC samplers. We demonstrate the utility of our recipe in "reinventing" previous stochastic gradient MCMC samplers, and in proposing our SGRHMC method. The efficiency and scalability of the SGRHMC method was shown on simulated data and a streaming Wikipedia analysis.

6.2 Stochastic Gradient MCMC Methods for Hidden Markov Models

Stochastic gradient based algorithms have proven crucial in numerous areas for scaling algorithms to large datasets. The key idea is to employ noisy estimates of the gradient based on minibatches of data, avoiding a costly gradient computation using the full dataset [148]. Assuming the data are i.i.d., the stochastic gradient is an unbiased estimate of the true gradient. In the context of Bayesian inference, such approaches have proven useful in scaling variational inference [67, 23, 22, 45] and Markov chain Monte Carlo (MCMC) [182, 121, 28, 37, 161]. For the latter, a primary focus has been on the influence of the stochastic gradient noise on the MCMC iterates; in contrast to many optimization-based procedures, it is non-trivial to show that the underlying (stochastic) dynamics maintain the correct stationary distribution in the presence of such noise. Significant headway has been made in developing such correct SG-MCMC procedures. For example, recently a recipe was proposed that translates the challenging problem of constructing efficient SG-MCMC algorithms into one of simply selecting two matrices [102]. Collectively, these algorithms have also shown great practical benefits and have gained significant traction.

A separate challenge, however, is the important and often overlooked question of whether such stochastic gradient techniques can be applied to massive amounts of *sequential* or otherwise *non-i.i.d.* data. In such cases, crucial dependencies must be broken to form the necessary minibatches. This question received some attention in the stochastic variational inference (SVI) algorithm of [45] for hidden Markov models (HMMs). In this work, we also focus in on HMMs as a simple example of a sequential data model, but turn our attention to SG-MCMC algorithms.

There are many existing algorithm to perform inference of the model parameters of an HMM including Monte Carlo methods [156], expectation-maximization [18], and variational algorithms [13]. All of these ideas operate by iterating between a *local update* for the latent states, followed by a *global update* of the model parameters. The local update is usually performed using the *forward-backward* algorithm that allows computation of any marginal, or pair-wise marginal distribution in time linear in the length of the sequence.

In the variational context, recent work has focused on scaling these *local-global* inference schemes to settings with a large number of replicates of short sequences [77, 69]. These methods utilize the fact that independent replicates of the observation sequence can be used to compute unbiased gradient estimates [77], and can be used to incrementally update sufficient statistics [69]. In contrast, the SVI algorithm of [45] examines how to deal with extremely long observation sequences. The algorithm heuristically breaks the dependence between observations and performs local updates on short subsequences of observations using a limited forward-backward algorithm. These existing methods suffer from a number of drawbacks. The variational approaches must use an approximate posterior distribution for both the state- and model-parameters, which may not be representative of the true distributions. The methods are also limited to conjugate prior distributions over the parameters, which can severely limit the expressiveness of the model. Finally, all of the methods discussed thus far are susceptible to becoming trapped in local modes during inference.

Unfortunately, existing SG-MCMC approaches cannot be adapted to HMMs by simply deploying the ideas of [45] within the MCMC context. The first challenge is that SG-MCMC methods rely on sampling continuous-valued parameter representations, whereas the HMM learning objective is typically specified in terms of the discrete-valued state sequence (local variables). To address this challenge, we consider an alternative approach to performing parameter inference for HMMs. Specifically, we work directly with the marginal likelihood of the observations sequence. Our algorithm then evaluates the marginal likelihood on a subsequence of observations.

The second challenge is in handling the dependencies between subsequences, in particular: i) the error introduced at the endpoints of a subsequence when considering the subsequence alone, and ii) the fact that subsequences are mutually correlated. We address both of these challenges by capitalizing on the memory decay of the Markovian structure underlying the data generating process. As in [45], we introduce short buffers around subsequences of observations that mitigate the error incurred. However, in contrast to the heuristic buffering schemes proposed for SVI-HMM, we provide a theoretically justified approach that estimates the buffer length. We also enforce that subsequences are separated by a minimum gap that ensures computations with them are uncorrelated. Collectively, the contributions of this paper make a sizable step towards general purpose SG-MCMC algorithms for sequential data.

Our synthetic data experiments investigate the impact of errors introduced via considering subchains naively, and how our buffering scheme can alleviate these issues. We then explore the benefits of our method over SVI-HMM. Finally, we provide an exploration of the computational gains over batch MCMC methods in an ion channel segmentation task. Here, our SG-MCMC algorithm provides good performance in segmenting the ion channel data about $1,000$ times faster than the batch MCMC method.

6.2.1 Background

Hidden Markov Models

Hidden Markov models (HMMs) are a class of discrete-time stochastic processes consisting of (i) latent discrte-valued observations $x_t \in \{1, \dots, K\}$ generated by a Markov chain and (ii) corresponding observations y_t generated from distributions determined by the latent states x_t. Specifically, for an observation sequence $\mathbf{y} = (y_1, \cdots, y_T)$ and latent state sequence $\mathbf{x} = (x_0, \cdots, x_T)$, the joint distribution factorizes as

$$p(\mathbf{x}, \mathbf{y}) = \pi_0(x_0) \prod_{t=1}^{T} p(x_t | x_{t-1}, A) \cdot p(y_t | x_t, \phi), \tag{6.18}$$

where A is the Markov transition matrix such that $A_{i,j} = \Pr(x_t = i | x_{t-1} = j)$, $\{\phi_k\}_{k=1}^{K}$ are the emission parameters, and $\pi_0 = p(x_0)$ is the initial state distribution. We denote the parameters of an HMM as $\theta = \{A, \phi\}$ and do not focus on performing inference on π_0.

Traditionally, expectation-maximization, variational inference, and Markov chain Monte Carlo are used to perform inference over θ [156, 13]. These algorithms rely on the well-known *forward-backward algorithm* to compute the marginal, $p(x_t | y_{1:T})$, and pairwise-marginal, $p(x_t, x_{t+1} | y_{1:T})$, distributions. The algorithm works by recursively computing a sequence of forward messages $\alpha_t(x_t) = p(x_t|y_{1:t})$ and backwards messages $\beta_t(x_t) = p(y_{t+1:T}|x_t)$ which can then be used to compute the necessary marginals [13]. These marginals are then used to update or sample from the distribution of the model parameters.

These past algorithms have found wide-spread use in statistics and machine learning. However, as discussed in Sec. 5, an alternative formulation of the HMM can provide greater utility in developing an SG-MCMC approach. Marginalizing over $\mathbf{x}$, we obtain the marginal likelihood of an observed sequence:

$$p(\mathbf{y}|\theta) = \mathbf{1}^{\mathrm{T}}\, P(y_T)A \cdots P(y_1)A\, \boldsymbol{\pi}_0, \tag{6.19}$$

where $P(y_T)$ is a diagonal matrix with $P_{i,i}(y_t) = p(y_t|x_t = i, \phi_i)$; $\mathbf{1}^{\mathrm{T}}$ is a row vector of K ones; and $(\boldsymbol{\pi}_0)_i = \pi_0(x_0 = i)$. The *posterior distribution* of θ given $\mathbf{y} = y_{1:T}$ is then:

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta). \tag{6.20}$$

Working with the marginal likelihood and posterior alleviates the need to compute the marginals and pairwise marginals of x_t. As such, only the forward pass of the forward-backward algorithm is performed. Indeed, performing the matrix multiplications in E-q. (6.19) from right to left corresponds to computing the normalizing constants of the forward messages. Similarly, performing the matrix multiplies from left to right corresponds to unnormalized messages in belief propagation, [46]. Perhaps most importantly for the development of our SG-MCMC algorithm, the marginal likelihood does not involve alternately updating the local state variables, x_t, and the global model parameters θ. Instead, we need only explore a continuous space which will allow us to leverage gradient information to develop a computationally and statistically efficient algorithm. The major impediment to directly using Eq. (6.19) for SG-MCMC is mitigating the error introduced when only dealing with subsequences.

Stochastic Gradient MCMC for i.i.d. Data

We leverage the SG-MCMC framework of [102, 103] which found that any sampler using continuous Markov dynamics can be represented as follows. If we define the potential function $U(\theta) \propto -\ln p(\theta|\mathbf{y})$ as the negative log posterior, then any such sampler that preserves the posterior distribution has an update rule:

$$\begin{aligned}\theta^{(t+1)} \leftarrow &\theta^{(t)} - \epsilon_t \left[\left(\mathbf{D}(\theta^{(t)}) + \mathbf{Q}(\theta^{(t)})\right)\nabla U(\theta^{(t)}) + \Gamma(\theta^{(t)})\right] \\ &+ \mathcal{N}(0, \epsilon_t(2\mathbf{D}(\theta^{(t)}))),\end{aligned} \tag{6.21}$$

where $\Gamma_i(\theta) = \sum_{j=1}^{d} \frac{\partial}{\partial \theta_j}(\mathbf{D}_{i,j}(\theta) + \mathbf{Q}_{i,j}(\theta))$, $\mathbf{D}(\theta^{(t)})$ is any positive-definite matrix and $\mathbf{Q}(\theta^{(t)})$ is any skew-symmetric matrix.

For i.i.d. data, the posterior distribution can be written as $p(\theta|\mathbf{y}) \propto \prod_{\mathbf{y}\in\mathcal{S}} p(\mathbf{y}|\theta) \cdot p(\theta)$ and the potential as $U(\theta) = -\sum_{\mathbf{y}\in\mathcal{S}} \ln p(\mathbf{y}|\theta) - \log p(\theta)$. Stochastic gradient algorithms examine *independently sampled* data subsets $\widetilde{\mathcal{S}} \subset \mathcal{S}$ resulting in a noisy estimate of the potential function:

$$\widetilde{U}(\theta) = -\frac{|\mathcal{S}|}{|\widetilde{\mathcal{S}}|} \sum_{\mathbf{y}\in\widetilde{\mathcal{S}}} \log p(\mathbf{y}|\theta) - \log p(\theta); \quad \widetilde{\mathcal{S}} \subset \mathcal{S}. \tag{6.22}$$

The specific form of Eq. (6.22) implies that $\widetilde{U}(\theta)$ is an unbiased estimator of $U(\theta)$. As such, a gradient computed based on $\widetilde{U}(\theta)$—called a *stochastic gradient*—is a noisy, but unbiased estimator of the full-data gradient. The key question in many of the existing stochastic gradient MCMC algorithms is whether the noise injected by the stochastic gradient adversely affects the stationary distribution of the modified dynamics (using $\nabla\widetilde{U}(\theta)$ in place of $\nabla U(\theta)$). One way to analyze the impact of the stochastic gradient is to make use of the central limit theorem and assume

$$\nabla\widetilde{U}(\theta) = \nabla U(\theta) + \mathcal{N}(0, \mathbf{V}(\theta)). \tag{6.23}$$

Simply using $\nabla\widetilde{U}(\theta)$ in place of $\nabla U(\theta)$ results in an additional noise term: $\big(\mathbf{D}(\theta) + \mathbf{Q}(\theta)\big)\ \mathcal{N}(0, \mathbf{V}(\theta))^T$. Assuming we have an estimate $\widehat{\mathbf{B}}$ of the variance of this additional noise satisfying $2\mathbf{D}(\theta) - \epsilon\widehat{\mathbf{B}} \succeq 0$ (i.e., positive semidefinite), then we can attempt to account for the stochastic gradient noise by simulating

$$\begin{aligned}\theta^{(t+1)} \leftarrow & \theta^{(t)} - \epsilon_t \Big[\big(\mathbf{D}(\theta^{(t)}) + \mathbf{Q}(\theta^{(t)})\big)\nabla\widetilde{U}(\theta^{(t)}) + \Gamma(\theta^{(t)})\Big] \\ & + \mathcal{N}(0, \epsilon_t(2\mathbf{D}(\theta^{(t)}) - \epsilon_t\widehat{\mathbf{B}}^{(t)})).\end{aligned} \tag{6.24}$$

This is the **stochastic gradient** MCMC algorithm for i.i.d. data proposed by [102, 103]. See Alg. 8.

6.2.2 Stochastic gradient MCMC for HMMs

In order to apply the SG-MCMC methodology to HMMs we must address three problems. First, we must be able to compute the gradient of the marginal likelihood efficiently for large data sets. Instead of using minibatches of individual observations as in standard SG-MCMC, we take the minibatches to be subsequences of consecutive observations. The

Algorithm 7: SG-MCMC

```
initialize $\theta_0$ for $t = 0, 1, 2 \cdots, N_{iter}$ do
    for $i = 1 \cdots n$ do
        $\Gamma_i(\theta) = \sum_j \frac{\partial}{\partial \theta_j} (\mathbf{D}_{ij}(\theta) + \mathbf{Q}_{ij}(\theta))$
    end
    sample $\eta^{(t)} \sim \mathcal{N}(0, 2\epsilon_t \mathbf{D}(\theta^{(t)}) - \epsilon_t^2 \widehat{\mathbf{B}}^{(t)})$
    $\theta^{(t+1)} \leftarrow \theta^{(t)} + \eta^{(t)} - \epsilon_t \Big[ \big(\mathbf{D}(\theta^{(t)}) + \mathbf{Q}(\theta^{(t)})\big) \nabla \widetilde{U}(\theta^{(t)}) + \Gamma(\theta^{(t)}) \Big]$
end
```

second problem is mitigating the error introduced by breaking dependencies at the endpoints of the subsequences. Finally, we show that SG-MCMC admits the desired stationary distribution when using subsequences of observations.

Recall from Sec. 6.2.1 that the posterior under an HMM is given by Eqs. (6.19) and (6.20) as:

$$p(\theta|\mathbf{y}) \propto \mathbf{1}^{\mathrm{T}} \, P(y_T)A \cdots P(y_1)A \, \boldsymbol{\pi}_0 \cdot p(\theta), \tag{6.25}$$

and that the *potential function* $U(\theta) \propto -\ln p(\theta|\mathbf{y})$. Both evaluating and computing gradients of Eq. (6.19) is computationally expensive when T is massive, for instance in problems arising in genomics. Instead, we propose only performing computations on a small subsequence of the observations. Define the *subsequence* with *half-width* L centered at time $\tau \in \{L+1, \ldots, T-L-1\}$ as $\mathbf{y}_{\tau,L} = (y_{\tau-L}, \ldots, y_\tau, \ldots, y_{\tau+L})$. The overall subsequence length is $2L+1$. The contribution to the likelihood of $\mathbf{y}_{\tau,L}$ is denoted

$$P(\mathbf{y}_{\tau,L}) = P(y_{\tau+L})A \cdots P(y_{\tau-L})A. \tag{6.26}$$

In fact, given a partition of $y_{1:T}$ into non-overlapping subchains, $\mathcal{S} = \{y_{\tau,L}\}$, we rewrite Eq. (6.25) as

$$\begin{aligned} p(\theta|\mathbf{y}) &\propto \mathbf{1}^{\mathrm{T}} \prod_{\mathbf{y}_{\tau,L} \in \mathcal{S}} P(\mathbf{y}_{\tau,L}) \cdot \boldsymbol{\pi}_0 \cdot p(\theta) \\ &= \mathbf{q}_{\tau+L+1} P(\mathbf{y}_{\tau,L}) \boldsymbol{\pi}_{\tau-L-1} \cdot p(\theta) \end{aligned} \tag{6.27}$$

where $\mathbf{q}_{\tau+L+1} = p(y_{\tau+L+1}|x_{\tau+L})$ is the likelihood of the observations that occur after $\mathbf{y}_{\tau,L}$ given the value of the latent state at τ, and $\boldsymbol{\pi}_{\tau-L-1} = p(x_{\tau-L}|y_{1:\tau-L-1})$ is the *predictive* *distribution* of the latent state at τ given the observations that occur before $\mathbf{y}_{\tau,L}$. Notice that we do not actually need to instantiate the latent state variables $x_{\tau-L}$ and $x_{\tau+L+1}$ as $\mathbf{q}_{\tau+L+1}$ and $\boldsymbol{\pi}_{\tau-L-1}$ can be computed (in theory) via the forward-backward algorithm [143, 156]. The gradient of the log-posterior from Eq. (6.27) is then given by

$$\begin{aligned}\frac{\partial U(\theta)}{\partial \theta_i} &= -\frac{\partial \ln p(\mathbf{y}|\theta)}{\partial \theta_i} - \frac{\partial p(\theta)}{\partial \theta_i} \\ &= -\sum_{\mathbf{y}_{\tau,L} \in \mathcal{S}} \frac{\mathbf{q}_{\tau+L+1}^{\mathrm{T}} \dfrac{\partial P(\mathbf{y}_\tau)}{\partial \theta_i} \boldsymbol{\pi}_{\tau-L-1}}{\mathbf{q}_{\tau+L+1}^{\mathrm{T}} P(\mathbf{y}_\tau) \boldsymbol{\pi}_{\tau-L-1}} - \frac{\partial \ln p(\theta)}{\partial \theta_i},\end{aligned} \tag{6.28}$$

where the equality follows from the product rule (see the Supplement for complete derivation).

We could imagine using $\nabla U(\theta)$ from Eq. (6.28) in the update rule of Eq. (6.21) to generate sample values of θ. However, Eq. (6.28) is expensive to compute as $\mathbf{q}_{\tau+L+1}$ and $\boldsymbol{\pi}_{\tau-L-1}$ require touching all T observations. This is prohibitive when T is massive. We instead propose to compute noisy estimates of Eq. (6.28) using only individual or small collections of subsequences, akin to the stochastic gradient updates of Eqs. (6.22)-(6.24), but for our non-i.i.d. scenario.

Stochastic Gradient Calculation

In order to use subsequences to reduce the amount of computation when computing $\nabla U(\theta)$, we inevitably introduce error at the boundaries of the subsequences where dependencies between observations are broken, as shown in Fig. 6.4. In terms of Eq. (6.28), we do not have the exact values of $\mathbf{q}_{\tau+L+1}$ and $\pi_{\tau-L-1}$, which would be prohibitively expensive to compute, so we instead approximate these terms.

Gradient Computation with Subsequences Inspired by recent work on stochastic variational inference for HMMs [45], we introduce a *buffer* of length B on either end of each subsequence $(\mathbf{y}_{LB}, \mathbf{y}_{\tau,L}, \mathbf{y}_{RB})$ where

$$\mathbf{y}_{LB} = (y_{\tau-L-B}, \ldots, y_{\tau-L-1}),$$

and

$$\mathbf{y}_{RB} = (y_{\tau+L+1}, \ldots, y_{\tau+L+B}).$$

See Fig. 6.4 for a figurative demonstration. For an irreducible and aperiodic Markov chain, the buffer regions will render the observations within $\mathbf{y}_\tau$ and those outside the buffers

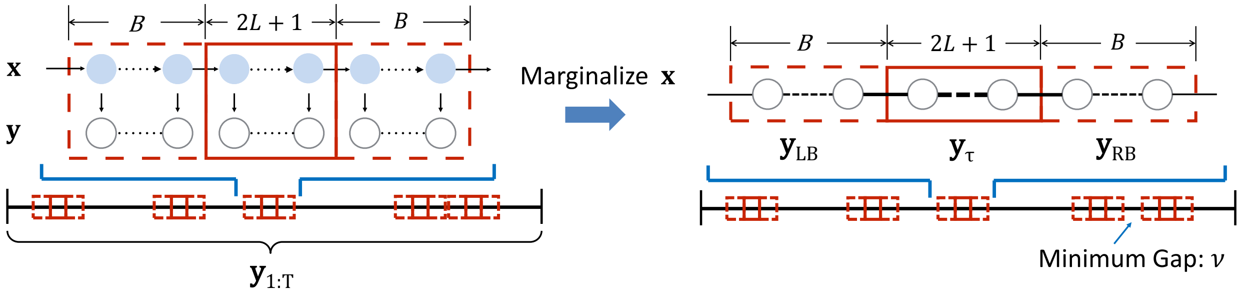


Figure 6.4: Diagram of subsequences, buffers, and subsequence sampling from full observation sequence. *Left:* The SVI method of [45] approximates stochastic gradients using subchains of length $2L + 1$ using the forward-backward algorithm performed on both the subchains and the associated buffer chains of length B. *Right:* Our propsoed SG-MCMC method uses a similar subsampling approach, however, i) the latent chain is never instantiated and ii) a minimum gap between consecutive subchains, $\mathbf{y}_{\tau,L}$, is used to ensure nearly uncorrelated subsequences. The thick black lines through the observables $\mathbf{y}$ represent all pairwise correlations between observations due to marginalization of $\mathbf{x}$. Correlation decays with distance enabling the segmentation of the of the chain into subsequences.

approximately independent. This lets us approximate the boundary terms in Eq. (6.28) as

$$\begin{aligned} \boldsymbol{\pi}_{\tau-L-1} &\approx \tilde{\boldsymbol{\pi}}_{\tau-L-1} = \underbrace{p(y_{\tau-L-1})A\cdots P(y_{\tau-L-B})A}_{P(\mathbf{y}_{LB})}\boldsymbol{\pi}_0 \\ \mathbf{q}^{\mathrm{T}}_{\tau+L+1} &\approx \tilde{\mathbf{q}}^{\mathrm{T}}_{\tau+L+1} = \mathbf{1}^{\mathrm{T}}\underbrace{P(y_{\tau+L+B})A\cdots P(y_{\tau+L+1})}_{P(\mathbf{y}_{RB})}. \end{aligned} \tag{6.29}$$

Notice that we plug in $\boldsymbol{\pi}_0$ and $\mathbf{1}^{\mathrm{T}}$ as the initial conditions for the buffers in Eq. (6.29). Though this introduces errors into the computations of $P(\mathbf{y}_{LB})$ and $P(\mathbf{y}_{RB})$, these errors will be nearly forgotten when processing observations in the subchain of interest $\mathbf{y}_\tau$ due to the mixing of the underlying Markov chain. We rewrite each term in Eq. (6.28) as

$$\frac{\mathbf{1}^{\mathrm{T}}P(\mathbf{y}_{RB})\dfrac{\partial P(\mathbf{y}_{\tau,L})}{\partial\theta}P(\mathbf{y}_{LB})\boldsymbol{\pi}_0}{\mathbf{1}^{\mathrm{T}}P(\mathbf{y}_{RB})P(\mathbf{y}_{\tau,L})P(\mathbf{y}_{LB})\boldsymbol{\pi}_0}. \tag{6.30}$$

In order to estimate $\nabla U(\theta)$ efficiently, we sample a collection of subsequences, $\tilde{S} = \{\mathbf{y}_{\tau,L}\}$ where $|\widetilde{S}|$ denotes the number of subchains. The τs are drawn randomly from $\{L+1,\dots,T-L-1\}$. We then use the following estimator of the full gradient

$$\frac{\partial\tilde{U}(\theta)}{\partial\theta_i} = -\frac{1}{p(\widetilde{\mathcal{S}})}\sum_{\mathbf{y}_{\tau,L}\in\tilde{S}}\frac{\mathbf{1}^{\mathrm{T}}P(\mathbf{y}_{RB})\dfrac{\partial P(\mathbf{y}_{\tau,L})}{\partial\theta_i}P(\mathbf{y}_{LB})\boldsymbol{\pi}_0}{\mathbf{1}^{\mathrm{T}}P(\mathbf{y}_{RB})P(\mathbf{y}_{\tau,L})P(\mathbf{y}_{LB})\boldsymbol{\pi}_0} - \frac{\partial\ln p(\theta)}{\partial\theta_i}, \tag{6.31}$$

where if we sample y_τ uniformly from $\mathcal{S}$, $p(\widetilde{\mathcal{S}}) = \dfrac{|\tilde{S}|L}{T}$, so that $\mathbb{E}\left[\nabla\tilde{U}(\theta)\right] = \nabla U(\theta)$ [58]. We note that Eq. (6.31) is computed in time $O(|\tilde{S}|LK^2)$. When $|\tilde{S}|L \ll T$ this results in significant computational speedups over batch inference algorithms.

A critical question that needs to be answered is how long the buffers should be? Though raised in previous work, only a heuristic solution was suggested [45]. We propose estimating the buffer length using the *Lyapunov exponent* of the *random dynamical system* specified by A and $P(\mathbf{y}_t)$. The Lyapunov exponent $\mathfrak{L}$ measures the evolution of the distance between vectors after applying the operator $(P(y_t)A)[\cdot]$ [10]. By generalizing the Perron–Frobenius theorem, all of the eigenvalues of the operator $(P(y_t)A)[\cdot]$ are less than 0, which implies that $\mathfrak{L} \leq 0$ [160]. The greater the absolute value of $\mathfrak{L}$, the faster the errors at the boundaries of the buffers decay, and the shorter the buffers need to be. Given an estimate of $\mathfrak{L}$, we set the buffer length as $B = \left\lceil \dfrac{1}{\mathfrak{L}}\ln\left(\dfrac{\delta}{\delta_0}\right)\right\rceil$ where $\delta \leq \delta_0$ are error tolerances. The method of calculating $\mathfrak{L}$ is described in the Supplement. Forthcoming work in the applied probability literature formalizes the validity of this approach.

Approximately Independent Minibatches When the subsequences used to estimate Eq. (6.31) are sampled naively, they will often overlap which diminishes the statistical efficiency of the estimator, requiring more subsequences to obtain accurate estimates. If we assume that the Markov chain of the latent state sequence is in equilibrium — a realistic assumption if T is huge — then we can leverage the memory decay of the Markov chain to encourage independent subsequences for use in the gradient estimator.

The mixing time of a Markov chain, denoted ν, is the number of steps needed until the chain is "close" to its stationary distribution $\boldsymbol{\pi}_0$ [160]. This implies that for $|t - t'| > \nu$, the corresponding x_t and $x_{t'}$ are approximately independent. Consequently, when $t < \tau - L - B$ or $t > \tau + L + B$, then y_t is approximately independent of $\mathbf{y}_{LB}$, $\mathbf{y}_{\tau,L}$, and $\mathbf{y}_{RB}$. Therefore, we can increase the statistical efficiency of $\nabla \tilde{U}(\theta)$ by sampling the $\mathbf{y}_{\tau,L}$s such that they are at least $2(L + B) + \nu$ time steps apart. Enforcing this non-overlapping structure results in $p(\widetilde{\mathcal{S}}) = \sum_{n=0}^{R-1} \frac{L}{T - n(L + 2B + 2L)}$ used in Eq. (6.31) to estimate $\nabla_\theta U(\theta)$. We estimate the mixing time $\nu = (1 - \hat{\lambda}_2)^{-1}$ where $\hat{\lambda}_2$ is the second largest eigenvalue of the current transition parameter iterate, $A^{(t)}$.

When sampling subsequences adhering to the mixing-time-dependent gap, each term in Eq. (6.31) is rendered approximately independent. Appealing to the central limit theorem we then have that

$$\frac{\partial \tilde{U}(\theta)}{\partial \theta_i} \approx \frac{\partial U(\theta)}{\partial \theta_i} + \mathcal{N}(0, V_i(\theta)). \tag{6.32}$$

As such, we can use the stationarity results from [102] and [103] to show that the proposed SG-MCMC for HMMs has the right stationary distribution in the small ϵ_t limit.

Incorporating Geometric Information

Eq. (6.21) serves as a general purpose algorithm that theoretically attains the correct stationary distribution for any $\mathbf{D}$ and $\mathbf{Q}$ matrices when the step size ϵ_t approaches zero. But in practice, we need to take into account numerical stability during numerical integrals. For example, when we are sampling from the probability simplex, previous work has shown that taking the curvature of the parameter space into account is important [182, 102]. Since our transition parameters live on the simplex, we likewise incorporate the geometry of the parameter space by constructing a *stochastic-gradient Riemannian MCMC* (SG-RMCMC) algorithm.

SG-RLD for transition parameters In order to sample the transition matrix A we note that the columns of A are constrained to lie on the probability simplex. To address these constraints, we use the expanded mean parametrization: $A = \frac{|\hat{A}_{i,j}|}{\sum_i |\hat{A}_{i,j}|}$, similar to what [121] used for topic modeling. Evaluating $\nabla U(\theta)$ in Eq. (6.28) for $\theta = \hat{A}_{i,j}$, using Eq. (6.26) yields:

$$\frac{\partial \widetilde{U}(\theta)}{\partial \hat{A}_{i,j}} = -\frac{1}{p(\widetilde{\mathcal{S}})} \sum_{\mathbf{y}_{\tau,L} \in \widetilde{\mathcal{S}}} \sum_{t=\tau-L}^{\tau+L} \frac{(\widetilde{\mathbf{q}}_{\tau+L+1})_i \, P_{i,i}(y_t) \, (\widetilde{\boldsymbol{\pi}}_{\tau-L-1})_j}{\widetilde{\mathbf{q}}^{\mathrm{T}}_{\tau+L} P(y_t) \hat{A} \widetilde{\boldsymbol{\pi}}_{\tau-L}}. \tag{6.33}$$

Here, $\tilde{\boldsymbol{\pi}}_{\tau-L-1}$ and $\tilde{\mathbf{q}}_{\tau+L+1}$ are computed on the left and right buffers, respectively, according to Eq. (6.29). The terms inside the sum in Eq. (6.33) are analogous to the pairwise marginals of the latent state in traditional HMM inference algorithms. A detailed derivation of this gradient can be found in the Supplement.

By leveraging the flexible SG-MCMC update rule of Eq. (6.24), we remove the dependency on $\hat{A}_{i,j}$ from the denominator of Eq. (6.33) by selecting $\mathbf{D} = \hat{A}$ and $\mathbf{Q} = \mathbf{0}$. This yields the following update:

$$\hat{A}^{(t+1)}_{i,j} \leftarrow \hat{A}^{(t)}_{i,j} - \epsilon_t \left[\hat{A}^{(t)}_{i,j} \nabla \widetilde{U}(\hat{A}^{(t)}_{i,j}, \phi) + I \right] + \mathcal{N}\left(0, \epsilon_t \left(2\hat{A}^{(t)}_{i,j} - \epsilon_t \widehat{\mathbf{B}}^{(t)}_{i,j}\right)\right) \tag{6.34}$$

where ϕ denotes all other model parameters. We note that this pre-conditioned gradient takes advantage of the local geometry of the parameter space by pre-multiplying by a metric tensor that arises from Eq. (6.24).

SG-RLD for emission parameters Similarly to the transition parameters, we sample the emission parameters $\{\phi_k : k = 1, \ldots, K\}$, by evaluating $\nabla \tilde{U}(\theta)$ in Eq. (6.28) for $\theta = \phi_k$ Using Eq. (6.26). This results in the gradient:

$$\frac{\partial \widetilde{U}(\theta)}{\partial \phi_k} = -\frac{1}{p(\widetilde{\mathcal{S}})} \sum_{\mathbf{y}_{\tau,L} \in \widetilde{\mathcal{S}}} \sum_{t=\tau-L}^{\tau+L} \frac{(\widetilde{\mathbf{q}}_{\tau+L+1})_k \, P_{k,k}(y_t) \, (\widetilde{\boldsymbol{\pi}}_{\tau-L-1})_k}{\widetilde{\mathbf{q}}^{\mathrm{T}}_{\tau+L+1} P(y_t) A \widetilde{\boldsymbol{\pi}}_{\tau-L-1}} \cdot \frac{\partial \ln P_{k,k}(y_t)}{\partial \phi_k}. \tag{6.35}$$

Again, $\tilde{\boldsymbol{\pi}}_{\tau-L-1}$ and $\tilde{\mathbf{q}}_{\tau+L+1}$ are computed on the left and right buffers, respectively, according to Eq. (6.29). Similarly to the transition parameters, we account for the geometry of the parameter space by specifying an appropriate $\mathbf{D}$ and $\mathbf{Q}$ in Eq. (6.24) which in general depends on the form of $p(y_t|\phi)$. For exponential family emission distributions we recommend using the inverse of the *Fisher information matrix* as $\mathbf{D}$ [5].

In this paper we will focus on Gaussian emission distributions in which case we have

$$\frac{\partial \ln P_{k,k}(y_t)}{\partial \mu_k} = \Sigma_k^{-1}(\mu_k - y_t) \tag{6.36}$$

Algorithm 8: SG-MCMC for HMM

```
initialize $A^{(0)}$ and $\phi_k^{(0)}$ for $n = 0, 1, 2 \cdots, N_{\text{iter}}$ do
    Periodically estimate the buffer length $B$ and the minimum subchain gap $\nu$ according to Sec. 6.2.2. Sample subchains $\tilde{S}$ of length $L$ from $p(\widetilde{S})$. for $s = 1 \cdots N_{\text{steps}}$ do
        Update $\hat{A}^{(s)}$ according to Eq. (6.33) and (6.34)
    end
    Calculate $\hat{A} = \frac{1}{N_{\text{steps}}} \sum_{s=1}^{N_{\text{steps}}} \hat{A}^{(s)}$. Set $A_{i,j} \leftarrow |\hat{A}_{i,j}| \Big/ \sum_i |\hat{A}_{i,j}|$ for $s = 1 \cdots N_{\text{steps}}$ do
        update $\phi^{(s)}$ according to Eqs. (6.35)– (6.39)
    end
    Set $\phi = \frac{1}{N_{\text{steps}}} \sum_{s=1}^{N_{\text{steps}}} \phi^{(s)}$.
end
```

$$\frac{\partial \ln P_{k,k}(y_t)}{\partial \Sigma_k} = \frac{1}{2} \Sigma_k^{-1} \left(\Sigma_k - (\mu_k - y_t)(\mu_k - y_t)^{\mathrm{T}} \right) \Sigma_k^{-1}. \tag{6.37}$$

We plug these values into the SG-MCMC update of Eq. (6.24) using $\mathbf{D} = \Sigma$ to account for the geometry of the parameter space and $\mathbf{Q} = \mathbf{0}$. This leads to the update equations:

$$\mu_k^{(t+1)} \leftarrow \mu_k^{(t)} - \epsilon_t \left[\Sigma_k^{(t)} \nabla_{\mu_k} \widetilde{U}(A, \phi_k^{(t)}) \right] + \mathcal{N}(0, \epsilon_t (2\Sigma_k^{(t)} - \epsilon_t \widehat{\mathbf{B}}_t)). \tag{6.38}$$

$$\begin{aligned} \Sigma_k^{(t+1)} \leftarrow \Sigma_k^{(t)} - \epsilon_t & \left[\Sigma_k^{(t)} \nabla_{\mu_k} \widetilde{U}(A, \phi_k^{(t)}) \Sigma_k^{(t)} + \Sigma_k^{(t)} \right] \\ & + \mathcal{N}(0, \epsilon_t (2\Sigma_k^{(t)} \otimes \Sigma_k^{(t)} - \epsilon_t \widehat{\mathbf{B}}^{(t)})). \end{aligned} \tag{6.39}$$

It is possible when using Eq. (6.39) to obtain a $\Sigma^{(t+1)}$ that is not positive definite. In this case we reject the update and set $\Sigma^{(t+1)} = \Sigma^{(t)}$.

We have presented a full SG-MCMC algorithm to perform inference in HMMs for massive sequences of data. In particular, we only require computations on collections of small subsequences and attain the desired stationary distribution by mitigating the errors incurred by these approximations. Finally, we have shown how to incorporate geometric information about the parameter space in order to increase the numerical robustness of the algorithm.

6.2.3 Experiments

We evaluate the performance of our proposed SG-MCMC algorithm for HMMs on both synthetic and real data. In the first two synthetic experiments, we recover two hard-to-capture dynamics and illustrate the trade off between the subchain length L and the number of subchains per minibatch $|\widetilde{\mathcal{S}}|$. We also illustrate the importance of the buffers. In the last synthetic experiment, we demonstrate the flexibility of our approach by comparing our SG-MCMC algorithm on HMM models using i) non-conjugate log-normal emissions, and ii) conjugate Gaussian emissions. (Recall that SVI-HMM only handles conjugate models.)

6.2.4 Synthetic Data

We first design two synthetic experiments in order to illustrate the trade off between the choice of subchain length L and the number of subchains per minibatch $|\widetilde{\mathcal{S}}|$. We also demonstrate the importance of the buffer in these two experiments. We fix the total number of observations used by the algorithm to $(L + B) \cdot |\widetilde{\mathcal{S}}|$ while varying L and $|\widetilde{\mathcal{S}}|$ and show how the performance of the algorithm is affected. Following [45], we create two synthetic datasets both with $T = 20$ million observations and $K = 8$ latent states.

The first data set, *diagonally dominant* (DD), illustrates the potential benefit of large $|\tilde{S}|$, the number of subchains per minibatch. The Markov chain heavily self-transitions so that most subchains contain redundant information with observations generated from the same latent state. Although transitions are rarely observed, the emission means are set to be distinct so that this example is likelihood-dominated and highly identifiable. See Fig. 6.5 (top left). Thus, with a fixed computational budget defined by $L + B$ and $|\tilde{S}|$, we expect large $|\widetilde{\mathcal{S}}|$ to be preferable to large L, covering more of the observation sequence and avoiding poor local modes arising from redundant information.

The second dataset we consider contains two *reversed cycles* (RC): the Markov chain strongly transitions from states $1 \rightarrow 2 \rightarrow 3 \rightarrow 1$ and $5 \rightarrow 7 \rightarrow 6 \rightarrow 5$ with a small probability of transition between cycles via bridge states 4 and 8. See Fig. 6.5 (top right). The emission means for the two cycles are very similar but occur in reverse order with respect to the transitions. The emission variance is larger, making states 1 and 5, 2 and 6, 3 and 7 undiscernible by themselves. Transition information in observing long enough dynamics is thus crucial to identify between states $1, 2, 3$ and $5, 6, 7$; therefore, we expect a large L to be imperative. See the the Supplement for details on generating both synthetic datasets.

We use a non-conjugate flat prior to demonstrate the flexibility of our algorithm. We initialize with a short run of k-means clustering to ensure that different states have different

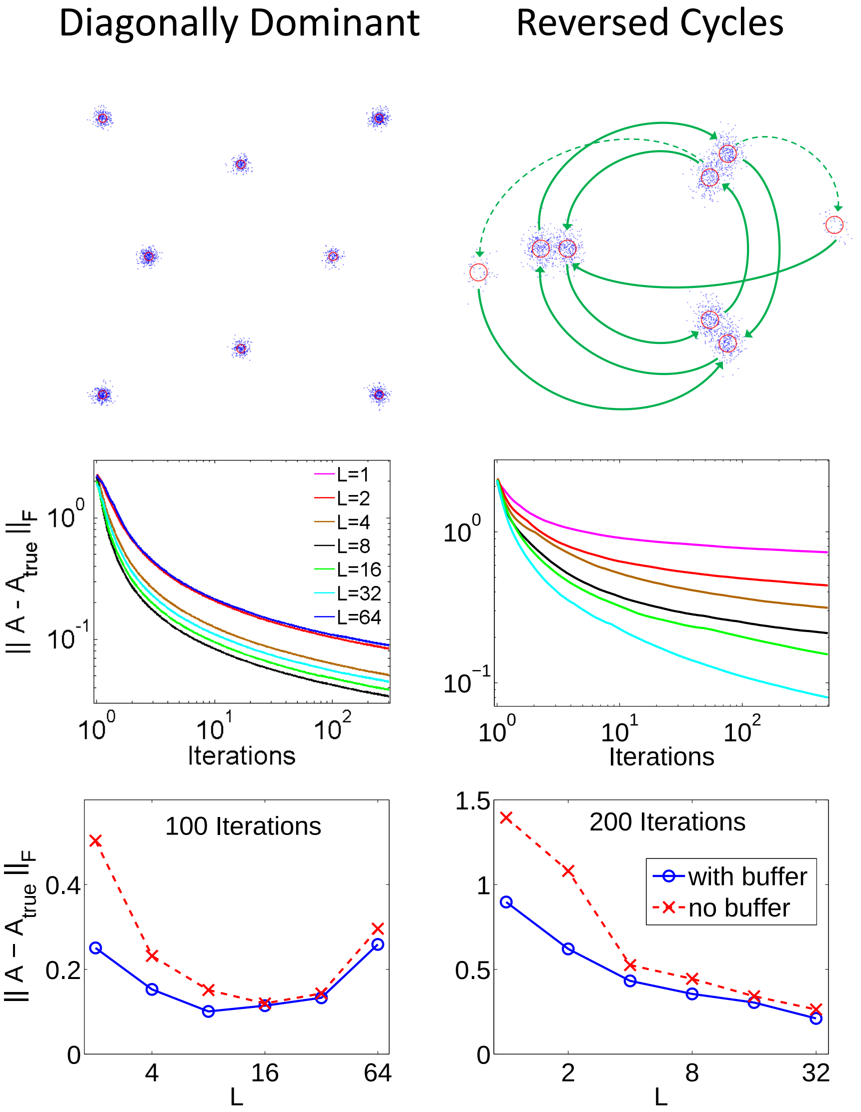
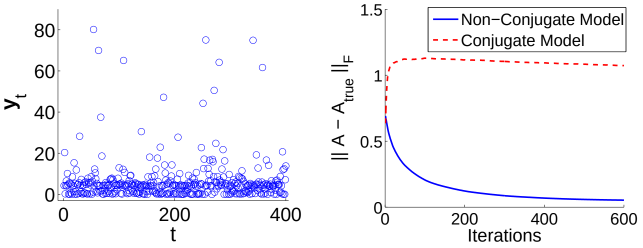


Figure 6.5: Synthetic experiments with hard-to-capture dynamics. Diagonally dominant (DD) (*left*) and reversed cycles (RC) (*right*) experiments. *First Row:* The emission distributions corresponding to 8 different states. Arrows in the RC case indicate the Markov transition structure with transition between bridge states as dashed arrows. *Second Row:* Loglog plot of error in transition parameter estimation versus iteration. *Third Row:* Comparison of SG-RLD with and without the buffers.



K	Non-Conjugate Emission	Conjugate Emission
1	-2.2679×10^4	-1.3172×10^5
2	$\mathbf{-1.9528 \times 10^4}$	-1.3081×10^5
3	-2.2005×10^4	-8.8227×10^4
4	-2.0376×10^4	$\mathbf{-7.5510 \times 10^4}$

Figure 6.6: Synthetic experiment with log-normal emission. We use the non-conjugate emission model on the synthetic data (*Top Left*) with two hidden states and log-normal emission and compare it against the conjugate model. We show the difference in convergence speed (*Top Right*) and log held out probability $\ln p(\mathbf{y}_{\text{test}}|\mathbf{y}_{\text{train}})$ on $2,000$ test data (*Right*).

emission parameters. In Fig. 6.5, we compare $||A - A_{\text{true}}||_F$, where A_{true} is the true transition matrix and A the mean estimate of the transition matrix under SG-MCMC with a specific setting of $(L + B)|\widetilde{S}|$. We see trends one would expect: the small L, large $|\widetilde{S}|$ settings achieve better performance for the DD example, but the opposite holds for RC, with $L = 2$ significantly underperforming.

As shown in the last row of Fig. 6.5, we examine the effect of using the buffer. We compare error in the estimate of the transition matrix after 100 and 200 iterations (for DD and RC experiments, respectively) with and without the buffer chain. Apart from the theoretical need for the buffers to derive the stochastic gradient MCMC algorithms, we see in practice the implications of the errors incurred at the edges of the subchains when they are not buffered. Long subsequences seem to mitigate the impact of these edge errors; however, as illustrated in the DD example, we prefer using many short subchains in persistent state examples, which are commonplace in real-world applications.

Non-conjugate emission distributions

We next demonstrate the benefit of our SG-RMCMC algorithm in being able to handle non-conjugate emissions. We simulate 2×10^5 observations from a 2 state HMM with log-normal emissions. Details of the parameter settings used to generate the data are in the Supplement. We evaluate the ability of two different HMM models in terms of parameter estimation and model selection accuracy.

The first HMM we consider uses log-normal emissions with non-conjugate normal pri-

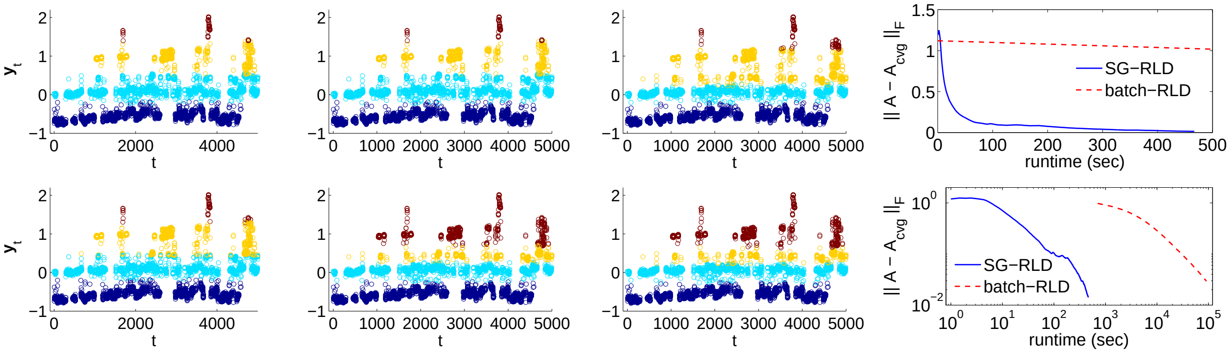


Figure 6.7: Inference of ion channel data *Top:* SG-RLD segmentation at runtimes: 44.05, 138.51, and 466.82 (sec). *Bottom:* Batch-RLD segmentation at runtimes: 716.19, 2124.43, and 7245.14 (sec). *Right:* Error decay of transition matrix estimates for SG-RLD and batch-RLD methods in original (top) and loglog (bottom) scales. A_{cvg} denotes the estimated transition parameters A after convergence. SG-RLD obtains plausible segmenations and accurate estimates of the transition matrix in a fraction of the time as a batch algorithm.

ors. The second model uses Gaussian emissions with a conjugate normal-inverse-Wishart prior. In Fig. 6.6 we show that the non-conjugate model obtains accurate estimates of the transition matrix in substantially fewer iterations than the conjugate model. Next, we demonstrate that efficiently handling non-conjugate models leads to improved model selection. Specificallly, we use SG-RLD to fit both the conjugate and non-conjugate HMMs described above with $K = 1, 2, 3, 4$ states and compute the marginal likelihood of the observations under each model. In the table of Fig. 6.6 we see that the non-conjugate model selects the right number of states (2), whereas the conjugate model selects a model with more states (4). The ability to use non-conjugate HMMs for truly massive data sets has not been feasible until this point and this experiment demonstrates its utility.

Ion Channel Recordings

We investigate the behavior of the SG-RLD sampler on ion channel recording data. In particular, we consider a 1MHz recording from [151] of a single alamethicin channel. This data was previously investigated in [120] and [169] using a complicated Bayesian nonparametric HMM. In that work, the authors downsample the data by a factor of 100 and only used $10,000$ and $2,000$ observations respectively due to the challenge of scaling computations to the full sequence. We subsample the time series by a factor of 50, resulting in $209,634$ observations, to reduce the strong autocorrelations present in the observations that are not captured well by a vanilla HMM. However, our algorithm would have no difficulty handling the full dataset. We further log-transform and normalize the observations to use Gaussian emission distributions.

We use a non-informative flat prior to analyze the ion channel data. In Fig. 6.7 we see that before the batch-RLD algorithm finishes a single iteration, the SG-RLD algorithm has already converged and generated a reasonable segmentation. With the converged estimation of the transition parameters A as reference, we calculated the speed of convergence of SG-RLD and batch-RLD algorithms and found that the SG-RLD is approximately $1,000$ times faster.

6.3 Discussion

We have developed an SG-MCMC algorithm to perform inference in HMM for massive observation sequences. The algorithm can be used with non-conjugate emission distributions and is thus applicable to modeling a variety of data. Also, the algorithm exactly samples from the posterior as opposed to variational approaches.

Developing the algorithm relied on three ingredients. First, we derived an efficient approach to estimate the gradient of the marginal likelihood of the HMM from only small subchains. Second, we developed a principled approach using buffers to mitigate the errors introduced when breaking the dependencies at the boundaries of the subchains. Unlike previous heuristic buffering schemes, our approach is theoretically justified using random dynamical systems. Last, we utilize sampling scheme based on the mixing time of the HMM to ensure subchains are approximately independent.

In future work we will extend these ideas to other models of dependent data, such as Markov random fields. Also, the ideas presented here are not limited to MCMC and could be used to develop more principled variational inference algorithms for dependent data.

BIBLIOGRAPHY

[1] S. Ahn, A. Korattikara, and M. Welling. Bayesian posterior sampling via stochastic gradient Fisher scoring. In *Proceedings of the 29th International Conference on Machine Learning (ICML'12)*, 2012.

[2] S. Ahn, B. Shahbaba, and M. Welling. Distributed stochastic gradient mcmc. In *Proceeding of 31st International Conference on Machine Learning (ICML'14)*, 2014.

[3] L. J. S. Allen. *An Introduction to Stochastic Processes with Applications to Biology*. Chapman & Hall/CRC, 2nd edition, 2010.

[4] B. Altaner. Foundations of stochastic thermodynamics. arXiv:1410.3983.

[5] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[6] P. W. Anderson. More is different: Broken symmetry and the nature of the hierarchical structure of science. *Science*, 177:393–396, 1972.

[7] L. Andrey. The rate of entropy change in non-hamiltonian systems. *Phys. Lett. A*, 111:45–46, 1985.

[8] C. Andrieu and J. Thoms. A tutorial on adaptive mcmc. *J. Stat. Comput.*, 18:343–373, 2008.

[9] P. Ao. Potential in stochastic differential equations: Novel construction. *J. Phys. A. Math. Gen.*, 37:L25–30, 2004.

[10] L. Arnold. *Random Dynamical Systems*. Springer, 1998.

[11] R. Bardenet, A. Doucet, and C. Holmes. Towards scaling up Markov chain Monte Carlo: An adaptive subsampling approach. In *Proceedings of the 30th International Conference on Machine Learning (ICML'14)*, 2014.

[12] R. Bardenet, A. Doucet, and C. Holmes. On Markov chain Monte Carlo methods for tall data. 2015.

[13] M. J. Beale. *Variational Algorithms for Approximate Bayesian Inference*. Ph.D. thesis, University College London, 2003.

[14] M. Betancourt. The fundamental incompatibility of scalable Hamiltonian Monte Carlo and naive data subsampling. In *Proceedings of the 31th International Conference on Machine Learning (ICML'15)*, 2015.

[15] J. Bierkens. Non-reversible Metropolis-Hastings. *J. Stat. Comput.*, pages 1–16, 2015.

[16] J. Bierkens, P. Fearnhead, and G. Roberts. The Zig-Zag process and super-efficient sampling for Bayesian analysis of big data. 2016.

[17] J. Bierkens and G. Roberts. A piecewise deterministic scaling limit of Lifted Metropolis-Hastings in the Curie-Weiss model. 2016.

[18] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

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

[20] N. Bou-Rabee and H. Owhadi. Long-run accuracy of variational integrators in the stochastic context. *SIAM J. Num. Anal.*, 48:278–297, 2010.

[21] A. Bouchard-Côté, S. J. Vollmer, and A. Doucet. The bouncy particle sampler: A non-reversible rejection-free Markov chain Monte Carlo method. 2016.

[22] T. Broderick, N. Boyd, A. Wibisono, A. C. Wilson, and M. I. Jordan. Streaming variational Bayes. In *Advances in Neural Information Processing Systems*, 2013.

[23] M. Bryant and E. B. Sudderth. Truly nonparametric online variational inference for hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems*, 2012.

[24] H. B. Callen. *Thermodynamics and an Introduction to Thermostatistics*. Wiley, 2nd edition, 1985.

[25] M. Campisi. On the mechanical foundations of thermodynamics: The generalized Helmholtz theorem. *Studies in History and Philosophy of Modern Physics*, 36:275–290.

[26] C. Chen, N. Ding, and L. Carin. On the convergence of stochastic gradient mcmc algorithms with high-order integrators. In *Advances in Neural Information Processing Systems 28*, pages 2278–2286. 2015.

[27] F. Chen, L. Lovász, and I. Pak. Lifting Markov chains to speed up mixing. In *Proceedings of the 31st annual ACM STOC*, pages 275–281. 1999.

[28] T. Chen, E. B. Fox, and C. Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *Proceeding of 31st International Conference on Machine Learning (ICML'14)*, 2014.

[29] T.-L. Chen and C.-R. Hwang. Accelerating reversible Markov chains. *Statistics & Probability Letters*, 83(9):1956–1962, 2013.

[30] W. Chen and M. J. Ward. The stability and dynamics of localized spot patterns in the two-dimensional Gray-Scott model. *SIAM J. Appl. Dyn. Syst.*, 10:582–666, 2011.

[31] S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

[32] R. T. Cox. Brownian motion in the theory of irreversible processes. *Rev. Mod. Phys.*, 24:312–320, 1952.

[33] G. Crooks. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Phys. Rev. E.*, 60:2721–2726, 1999.

[34] A. Dembo and J.-D. Deuschel. Markovian perturbation, response and fluctuation dissipation theorem. *Ann. Inst. H. Poincar Probab. Statist.*, 46:822–852, 2010.

[35] J. D. Deuschel and D. W. Stroock. *Large Deviations*. Amer. Math. Soc., 2001.

[36] P. Diaconis, S. Holmes, and R. M. Neal. Analysis of a nonreversible Markov chain sampler. *Ann. Appl. Probab.*, 10:726–752, 2000.

[37] 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 (NIPS'14)*. 2014.

[38] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216 – 222, 1987.

[39] A. B. Duncan, T. Lelièvre, and G. A. Pavliotis. Variance reduction using nonreversible Langevin samplers. *Journal of Statistical Physics*, 163(3):457–491, 2016.

[40] A. Durmus, G. O. Roberts, G. Vilmart, and K. C. Zygalakis. Fast Langevin based algorithm for MCMC in high dimensions. 2016.

[41] M. Esposito. Stochastic thermodynamics under coarse graining. *Phys. Rev. E*, 85:041125, 2012.

[42] Y. Fang, J. M. Sanz-Serna, , and R. D. Skeel. Compressible generalized hybrid Monte Carlo. *J. Chem. Phys.*, 140:174108, 2014.

[43] W. Feller. *Introduction to Probability Theory and its Applications*. John Wiley & Sons, 1950.

[44] J.-F. Feng. The hydrodynamic limit for the reaction diffusion equationm — an approach in terms of the gpv method. *J. Theoret. Prob.*, 9:285–299, 1996.

[45] N. J. Foti, J. Xu, D. Laird, and E. B. Fox. Stochastic variational inference for hidden Markov models. In *Advances in Neural Information Processing Systems*, 2014.

[46] E.B. Fox. *Bayesian Nonparametric Learning of Complex Dynamical Phenomena*. Ph.D. thesis, MIT, Cambridge, MA, 2009.

[47] R. F. Fox. Gaussian stochastic processes in physics. *Phys. Rep.*, 48:180–283, 1978.

[48] M. I. Freidlin and A. D. Wentzell. *Random Perturbations of Dynamical Systems*. Springer-Verlag, 3rd edition, 2012.

[49] G. Gallavotti. *Statistical Mechanics: A Short Treatise*. Springer, 1999.

[50] G. Gallavotti, W. L. Reiter, and J. Yngvason. *Boltzmanns Legacy*. Eur. Math. Soc. Pub., 2007.

[51] C. W. Gardiner. *Handbook of Stochastic Methods*. Springer, 4th edition, 2009.

[52] H. Ge and D.-Q. Jiang. The transient fluctuation theorem of sample entropy production for general stochastic processes. *J. Phys. A: Math. Theor.*, 40:F713–723, 2007.

[53] H. Ge and H. Qian. Maximum entropy principle, equal probability *a priori*, and gibbs paradox. arXiv:1105.4118.

[54] H. Ge, M. Qian, and H. Qian. Stochastic theory of nonequilibrium steady states (Part II): Applications in chemical biophysics. *Phys. Rep.*, 510:87–118, 2012.

[55] A. Gelman, J. B. Carhn, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, 2004.

[56] C. J. Geyer. Practical markov chain monte carlo. *Statist. Sci.*, 7:473–483, 1992.

[57] M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[58] P. K. Gopalan, S. Gerrish, M. Freedman, D. M. Blei, and D. M. Mimno. Scalable inference of overlapping communities. In *Advances in Neural Information Processing Systems*, pages 2249–2257. 2012.

[59] J. Grasman and O. A. van Herwaarden. *Asymptotic Methods for the Fokker-Planck Equation and the Exit Problem in Applications*. Springer, 1999.

[60] M.-Z. Guo, G. C. Papanicolaou, and S. R. S. Varadhan. Nonlinear diffusion limit for a system with nearest neighbor interactions. *Comm. Math. Phys.*, 118:31–59, 1988.

[61] P. Gustafson. A guided walk Metropolis algorithm. *Statistics and Computing*, 8(4):357–364, 1998.

[62] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402:C47–52, 1999.

[63] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:pp 97–109, 1970.

[64] T. Hatano and S.-I. Sasa. Steady-state thermodynamics of Langevin systems. *Phys. Rev. Lett.*, 86:3463–3466, 2001.

[65] T. L. Hill. *Thermodynamics of Small Systems*. Dover, 1994.

[66] M. D. Hoffman and A. Gelman. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *arXiv*, 1111.4246, 2011.

[67] M.D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Maching Learning Research*, 14(1):1303–1347, May 2013.

[68] A. M. Horowitz. A generalized guided Monte Carlo algorithm. *Physics Letters B*, 268(2):247 – 252, 1991.

[69] M. C. Hughes, W. Stephenson, and E. B. Sudderth. Scalable adaptation of state complexity for nonparametric hidden Markov models. In *Advances in Neural Information Processing Systems*, 2015.

[70] C.-R. Hwang, S.-Y. Hwang-Ma, and S.-J. Sheu. Accelerating Gaussian diffusions. *Ann. Appl. Probab.*, 3(3):897–913, 08 1993.

[71] C.-R. Hwang, S.-Y. Hwang-Ma, and S.-J. Sheu. Accelerating diffusions. *Ann. Appl. Probab.*, 15(2):1433–1444, 05 2005.

[72] Khinchin A. I. *Mathematical Foundations of Statistical Mechanics*. Dover, 1949.

[73] S. F. Jarner and G. O. Roberts. Convergence of heavy-tailed Monte Carlo Markov chain algorithms. *Scandinavian Journal of Statistics*, 34(4):781–815, 2007.

[74] C. Jarzynski. Nonequilibrium equality for free energy difference. *Phys. Rev. Lett.*, 78:2690–2693, 1997.

[75] E. T. Jaynes. *Probability Theory: The Logic of Science*. 2003.

[76] D.-Q. Jiang, M. Qian, and M.-P. Qian. *Mathematical Theory of Nonequilibrium Steady States*. Springer, 2004.

[77] M. J. Johnson and A. S. Willsky. Stochastic variational inference for Bayesian time series models. In *International Conference on Machine Learning*, 2014.

[78] A. N. Kolmogorov. Über die analytischen methoden in der wahrscheinlichkeitsrechnung. *Math. Ann.*, 104:415–458, 1931.

[79] A. N. Kolmogorov. Zur theorie der markoffschen ketten. *Math. Ann.*, 112:155–160, 1936.

[80] A. N. Kolmogorov. Three approaches to the quantitative definition of information. *Int. J. Comput. Math.*, 2:157–168, 1968.

[81] T. Komorowski, C. Landim, and S. Olla. *Fluctuations in Markov Processes – Time Symmetry and Martingale Approximation*. Springer, Berlin-Heidelberg-New York, 2012.

[82] A. Korattikara, Y. Chen, and M. Welling. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *Proceedings of the 30th International Conference on Machine Learning (ICML'14)*, 2014.

[83] M. Kot. *Elements of Mathematical Ecology*. Cambridge Univ. Press, 2001.

[84] S. C. Kou, Q. Zhou, and W. H. Wong. Discussion paper: Equi-energy sampler with applications in statistical inference and statistical mechanics. *The Annals of Statistics*, 34(4):1581–1619, 2006.

[85] J. Kurchan. Fluctuation theorem for stochastic dynamics. *J. Phys. A: Math. Gen.*, 31:3719–3729, 1998.

[86] T. G. Kurtz. *Approximation of Population Processes*. SIAM Pub., 1981.

[87] C. Kwon, P. Ao, and D. J. Thouless. Structure of stochastic dynamics near fixed points. *Proc. Natl. Acad. Sci. U. S. A.*, 102:13029–13033, 2005.

[88] J. S. W. Lamb and J. A. G. Roberts. Time-reversal symmetry in dynamical systems: A survey. *Physica D*, 112:1–39, 1998.

[89] S. Lan, V. Stathopoulos, B. Shahbaba, and M. Girolami. Lagrangian dynamical Monte Carlo. *arXiv*, 1211.3759, 2012.

[90] L. D. Landau and E. M. Lifshitz. *Statistical Physics*. 1980.

[91] M. Lax. Fluctuations from the nonequilibrium steady state. *Rev. Mod. Phys.*, 32:25–64, 1960.

[92] M. Lax. Classical noise iii: Nonlinear markoff processes. *Rev. Mod. Phys.*, 38:359–379, 1966.

[93] J. L. Lebowitz and H. Spohn. A gallavotti-cohen-type symmetry in the large deviation functional for stochastic dynamics. *J. Stat. Phys.*, 95:333–365, 1999.

[94] B. Leimkuhler, C. Matthews, and M. Tretyakov. On the long-time integration of stochastic gradient systems. *Proceedings of the Royal Society A*, 470:20140120, 2014.

[95] F. Lesmes, D. Hochberg, F. Morán, and J. Pérez-Mercader. Noise-controlled self-replicating patterns. *Phys. Rev. Lett.*, 91:238301, 2003.

[96] D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2008.

[97] G. N. Lewis. A new principle of equilibrium. *Proc. Natl. Acad. Sci. U. S. A.*, 11:179–183, 1925.

[98] T. Li and P. Zhou. Construction of the landscape for multi-stable systems: Potential landscape, quasi-potential, A-type integral and beyond. *J. Chem. Phys*, 144:p094109, 2016.

[99] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, 2004.

[100] A. J. Lotka. *Elements of Physical Biology*. Williams & Wilkins, 1925.

[101] Y. Ma, Q. Tan, R. Yuan, B. Yuan, and P. Ao. Potential function in a continuous dissipative chaotic system: Decomposition scheme and role of strange attractor. *Int. J. Bifur. Chaos*, 24:1450015, 2014.

[102] Y.-A Ma, T. Chen, and E. Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems 28*, pages 2899–2907. 2015.

[103] Y.-A. Ma, E. B. Fox, T. Chen, and L. Wu. A unifying framework for devising efficient and irreversible MCMC samplers. arXiv:1608.05973, 2016.

[104] Y.-A. Ma and H. Qian. A thermodynamic theory of ecology: Helmholtz theorem for Lotka-Volterra equation, extended conservation law, and stochastic predator-prey dynamics. *Proc. R. Soc. A*, 471:20150456, 2015.

[105] Y.-A Ma and H. Qian. Universal ideal behavior and macroscopic work relation of linear irreversible stochastic thermodynamics. *New Journal of Physics*, 17(6):065013, 2015.

[106] D. Mandal and C. Jarzynski. Work and information processing in a solvable model of Maxwell's demon. *Proc. Natl. Acad. Sci. U. S. A.*, 109:11641–11645, 2012.

[107] M. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and Teller. E. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21, 1953.

[108] C. B. Muratov, E. Vanden-Eijnden, and W. E. Noise-controlled self-replicating patterns. *Proc. Natl. Acad. Sci. U. S. A.*, 104:702–707, 2007.

[109] J. D. Murray. *Mathematical Biology I: An Introduction*. Springer, 3rd edition, 2002.

[110] R. M. Neal. Bayesian learning via stochastic dynamics. In *Advances in Neural Information Processing Systems 5 (NIPS'93)*, pages 475–482, 1993.

[111] R. M. Neal. *Bayesian Learning for Neural Networks*. Springer, 1996.

[112] R. M. Neal. Improving asymptotic variance of MCMC estimators: Non-reversible chains are better. 2004.

[113] R. M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 54:113–162, 2010.

[114] R. M. Noyes. Entropy of mixing of interconvertible species. some remarks on the gibbs paradox. *J. Chem. Phys.*, 34, 1961.

[115] L. Onsager. Reciprocal relations in irreversible processes. I. *Phys. Rev.*, 37:405–426, 1931.

[116] L. Onsager and S. Machlup. Fluctuations and irreversible process. II. systems with kinetic energy. *Phys. Rev.*, 91:1512–1515, 1953.

[117] L. Onsager and S. Machlup. Fluctuations and irreversible processes. *Phys. Rev.*, 91:1505–1512, 1953.

[118] M. Osterfield, X. Du, T. Schüpbach, E. Wieschaus, and S.Y. Shvartsman. Three-dimensional epithelial morphogenesis in the developing Drosophila egg. *Dev. Cell*, 24:400–410, 2013.

[119] M. Ottobre, N. S. Pillai, F. J. Pinski, and A. M. Stuart. A function space HMC algorithm with second order Langevin diffusion limit. *Bernoulli*, 22(1):60–106, 02 2016.

[120] K. Palla, D. A. Knowles, and Z. Ghahramani. A reversible infinite HMM using normalised random measures. In *International Conference on Machine Learning*, 2014.

[121] S. Patterson and Y. W. Teh. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems 26 (NIPS'13)*. 2013.

[122] W. Pauli. *Pauli Lectures on Physics: Thermodynamics and the Kinetic Theory of Gas*, volume 3. MIT Press, 1973.

[123] G. A. Pavliotis. *Stochastic Processes and Applications*. Springer, 2014.

[124] J. E. Pearson. Complex patterns in a simple system. *Science*, 261:189–192, 1993.

[125] D. M. Pfund, L. L. Lee, and H. D. Cochran. Chemical potential prediction in realistic fluid models with scaled particle theory. *Int. J. Thermophys.*, 11, 1990.

[126] M. Planck. *Treatise on Thermodynamics*. Dover, 1969.

[127] M. Polettini. Of dice and men. subjective priors, gauge invariance, and nonequilibrium thermodynamics. arXiv:1307.2057.

[128] M. Polettini. Nonequilibrium thermodynamics as a gauge theory. *EPL*, 97:30003, 2012.

[129] S. Pressé, K. Ghosh, J. Lee, and K. A. Dill. Principles of maximum entropy and maximum caliber in statistical physics. *Rev. Mod. Phys.*, 85:1115–1141, 2013.

[130] I. Prigogine. *Etude thermodynamique des phénomines irreversibles*. Desoer, 1947.

[131] H. Qian. Mathematical formalism for isothermal linear irreversibility. *Proc. Roy. Soc. A: Math. Phys. Engr. Sci.*, 457:1645–1655, 2001.

[132] H. Qian. Nonequilibrium steady-state circulation and heat dissipation functional. *Phys. Rev. E*, 64:022101, 2001.

[133] H. Qian. Nonlinear stochastic dynamics of mesoscopic homogeneous biochemical reaction systems - an analytical theory. *Nonlinearity*, 24:R19–R49, 2011.

[134] H. Qian. A decomposition of irreversible diffusion processes without detailed balance. *J. Math. Phys.*, 54:053302, 2013.

[135] H. Qian. The zeroth law of thermodynamics and volume-preserving conservative system in equilibrium with stochastic damping. *Phys. Lett. A*, 378:609–616, 2014.

[136] H. Qian. Thermodynamics of the general diffusion process: Equilibrium supercurrent and nonequilibrium driven circulation with dissipation. *Eur. Phys. J.*, 224:781–799, 2015.

[137] H. Qian and H. Ge. Analytical mechanics in stochastic dynamics: most probable path, large-deviation rate function and Hamilton-Jacobi equation. *Int. J. Mod. Phys. B*, 26:1230012, 2012.

[138] H. Qian and J. J. Hopfield. Entropy-enthalpy compensation: Perturbation and relaxation in thermodynamic systems. *J. Chem. Phys.*, 105:9292–9296, 1996.

[139] H. Qian, M. Qian, and X. Tang. Thermodynamics of the general diffusion process: Time-reversibility and entropy production. *Journal of Statistical Physics*, 107:1129, 2002.

[140] H. Qian and S. Roy. An information theoretical analysis of kinase activated phosphorylation dephosphorylation cycle. *IEEE Trans. NanoBiosci.*, 11:289–295, 2012.

[141] Hong Qian, Ping Ao, Yuhai Tu, and Jin Wang. A framework towards understanding mesoscopic phenomena: Emergent unpredictability, symmetry breaking and dynamics across scales. *Chem. Phys. Lett.*, 665:153–161, 2016.

[142] M.-P. Qian, M. Qian, and G.-L. Gong. The reversibility and the entropy production of Markov processes. *Contemp. Math.*, 118:255–261, 1991.

[143] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286, 1989.

[144] J. Raspopovic, L. Marcon, L. Russo, and J. Sharpe. Digit patterning is controlled by a Bmp-Sox9-Wnt Turing network modulated by morphogen gradients. *Science*, 345:566–570, 2014.

[145] L. Rey-Bellet and K. Spiliopoulos. Irreversible Langevin samplers and variance reduction: A large deviations approach. *Nonlinearity*, 28(7):2081, 2015.

[146] H. Risken. *The Fokker-Planck Equation, Methods of Solution and Applications*. Springer, 1996.

[147] H. Risken and T. Frank. *The Fokker-Planck Equation: Methods of Solutions and Applications*. Springer, 1996.

[148] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 09 1951.

[149] C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2nd edition, 2004.

[150] G. O. Roberts and O. Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology And Computing In Applied Probability*, 4:337–357, 2002.

[151] J. K. Rosenstein, S. Ramakrishnan, J. Roseman, and Shepard K. L. Single ion channel recordings with CMOS-anchored lipid membranes. *Nano Letters*, 13(6):2682–2686, 2013.

[152] R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th International Conference on Machine Learning (ICML'08)*, pages 880–887, 2008.

[153] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems 20 (NIPS'08)*, pages 1257–1264, 2008.

[154] M. Santillán and H. Qian. Irreversible thermodynamics in multiscale stochastic dynamical systems. *Phys. Rev. E*, 83:041130, 2011.

[155] M. Scott, F. J. Poulin, and H. Tang. Approximating intrinsic noise in continuous multispecies models. *Proc. R. Soc. A*, 467:718–737, 2011.

[156] S. L. Scott. Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, 97(457):337–351, 03 2002.

[157] U. Seifert. Entropy production along a stochastic trajectory and an integral fluctuation theorem. *Phys. Rev. Lett.*, 95:040602, 2005.

[158] U. Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Rep. Prog. Phys.*, 75:126001, 2012.

[159] K. Sekimoto. *Stochastic Energetics*. Springer, 2010.

[160] E. Seneta. *Non-negative matrices and Markov chains*. Springer Science & Business Media, 2006.

[161] X. Shang, Z. Zhu, B. Leimkuhler, and A. Storkey. Covariance-controlled adaptive Langevin thermostat for large-scale Bayesian sampling. In *Advances in Neural Information Processing Systems 28 (NIPS'15)*. 2015.

[162] J. Shi, T. Chen, R. Yuan, B. Yuan, and P. Ao. Relation of a new interpretation of stochastic differential equations to Itô process. *Journal of Statistical Physics*, 148(3):579–590, 2012.

[163] S. Y. Shvartsman, C. B. Muratov, and D. A. Lauffenburger. Modeling and computational analysis of EGF receptor-mediated cell communication in Drosophila oogenesis. *Development*, 129:2577–2589, 2002.

[164] I. Sutskever, J. Martens, G. E. Dahl, and G. E. Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning (ICML'13)*, volume 28, pages 1139–1147, May 2013.

[165] H. Tak, X.-L. Meng, and D. A. van Dyk. A repulsive-attractive Metropolis algorithm for multimodality. 2016.

[166] R. C. Tolman and P. C. Fine. On the irreversible production of entropy. *Rev. Mod. Phys.*, 20:51–77, 1948.

[167] N. Tompkins, N. Li, L. Russo, C. Girabawe, M. Heymann, G. B. Ermentrout, I. R. Epstein, and S. Fraden. Testing turings theory of morphogenesis in chemical cells. *Proc. Natl. Acad. Sci. U. S. A.*, 111:4397–4402, 2014.

[168] S. Toyabe, T. Sagawa, M. Ueda, E. Muneyuki, and M. Sano. Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality. *Nature Phys.*, 6:988–992, 2010.

[169] N. Tripuraneni, S. Gu, H. Ge, and Z. Ghahramani. Particle Gibbs for infinite hidden Markov Models. In *Advances in Neural Information Processing Systems*, pages 2386–2394, 2015.

[170] M. Tuckerman, Y. Liu, G. Ciccotti, and G. Martyna. Non-Hamiltonian molecular dynamics: Generalizing Hamiltonian phase space principles to non-Hamiltonian systems. *J. Chem. Phys.*, 115:1678–1702, 2001.

[171] A. M. Turing. The chemical basis of morphogenesis. *Philos. Trans. R. Soc. Lond.*, 237:37–72, 1952.

[172] K. S. Turitsyn, M. Chertkov, and M. Vucelja. Irreversible Monte Carlo algorithms for efficient sampling. *Physica D: Nonlinear Phenomena*, 240(4–5):410–414, 2011.

[173] C. van den Broeck and M. Esposito. Ensemble and trajectory thermodynamics: A brief introduction. *Physica A*, 418:6–16, 2015.

[174] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry*. North Holland, 2nd edition.

[175] C. Villani. *Hypocoercivity*. American Mathematical Soc., 2009.

[176] M. Vucelja. Lifting – a nonreversible Markov chain Monte Carlo algorithm. 2015.

[177] J. Wang, L. Xu, and E. Wang. Potential landscape and flux framework of nonequilibrium networks: Robustness, dissipation, and coherence of biochemical oscillations. *Proc. Natl. Acad. Sci. USA*, 105:12271–12276, 2008.

[178] M. C. Wang and G. E. Uhlenbeck. On the Theory of the Brownian Motion II. *Reviews of Modern Physics*, 17(2-3):323, 1945.

[179] Z. Wang, S. Mohamed, and D. Nando. Adaptive Hamiltonian and Riemann manifold Monte Carlo. In *Proceedings of the 30th International Conference on Machine Learning (ICML'13)*, volume 28, pages 1462–1470. JMLR Workshop and Conference Proceedings, May 2013.

[180] N. Wax. *Selected Papers on Noise and Stochastic Processes*. 1954.

[181] R. Wegscheider. Über simultane gleichgewichte und die beziehungen zwischen thermodynamik und reaktionskinetik homogener systeme. *Zeitschrift für Physikalische Chemie*, 39:257–303, 1902.

[182] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML'11)*, pages 681–688, June 2011.

[183] N. Wiener. *Cybernetics: or the Control and Communication in the Animal and the Machine*. MIT Press, 1948.

[184] T. Xifara, C. Sherlock, S. Livingstone, S. Byrne, and M. Girolami. Langevin diffusions and the Metropolis-adjusted Langevin algorithm. *Statistics and Probability Letters*, 91:14–19, 2014.

[185] L. Xu, F. Zhang, K. Zhang, E. Wang, and J. Wang. The potential and flux landscape theory of ecology. *PLoS ONE*, 9:e86746, 2014.

[186] L. Yin and P. Ao. Existence and construction of dynamical potential in nonequilibrium processes without detailed balance. *Journal of Physics A: Mathematical and General*, 39(27):8593, 2006.

[187] R. Yuan, Y. Ma, B. Yuan, and P. Ao. Lyapunov function as potential function: A dynamical equivalence. *Chin. Phys. B*, 23:010505, 2013.

[188] F. Zhang, L. Xu, K. Zhang, E. Wang, , and J. Wang. The potential and flux landscape theory of evolution. *J Chem Phys.*, 137:065102, 2012.

[189] X.-J. Zhang, H. Qian, and M. Qian. Stochastic theory of nonequilibrium steady states and its applications (Part I). *Phys. Rep.*, 510:1–86, 2012.

[190] W. Q. Zhu. Nonlinear stochastic dynamics and control in Hamiltonian formulation. *Trans. A.S.M.E.*, 59:230–248.

[191] R. Zwanzig. *Nonequilibrium Statistical Mechanics*. Oxford University Press, 2001.