Efficient Sampling Methods of, by, and for Stochastic Dynamical Systems

by

Benjamin Jiahong Zhang

S.M., Massachusetts Institute of Technology (2017)
B.S., B.A., University of California, Berkeley (2015)

Submitted to the Department of Aeronautics and Astronautics
and the Center for Computational Science and Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computational Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2022

Author .....
Department of Aeronautics and Astronautics
Center for Computational Science and Engineering
January 13, 2022

Certified by .....
Youssef M. Marzouk
Professor of Aeronautics and Astronautics
Thesis Supervisor

Certified by .....
Tuhin Sahai
Technical Fellow, Raytheon Technologies Research Center
Thesis Committee Member

Certified by .....
Themistoklis Sapsis
Associate Professor of Mechanical and Ocean Engineering
Thesis Committee Member

Certified by .....
Konstantinos Spiliopoulos
Associate Professor of Mathematics and Statistics, Boston University
Thesis Committee Member

Accepted by .....
Jonathan P. How
R. C. Maclaurin Professor of Aeronautics and Astronautics
Chair, Graduate Program Committee

Accepted by .....
Nicolas G. Hadjiconstantinou
Professor of Mechanical Engineering
Co-Director, Center for Computational Science and Engineering

Efficient Sampling Methods of, by, and for Stochastic Dynamical Systems

by

Benjamin Jiahong Zhang

Submitted to the Department of Aeronautics and Astronautics
and the Center for Computational Science and Engineering
on January 13, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computational Science and Engineering

Abstract

This thesis presents new methodologies that lie at the intersection of computational statistics and computational dynamics. Stochastic differential equations (SDEs) are used to model a variety of physical systems, and computing expectations over marginal distributions of SDEs is important for the analysis of such systems. In particular, quantifying the probabilities of rare events in SDEs—and elucidating the mechanisms by which these events occur—are critical to the design and safe operation of engineered systems.

In the first part of the thesis, we use data-driven tools for dynamical systems to create methods for efficient *rare event simulation in nonlinear SDEs*. Our approach exploits the relationship between the stochastic Koopman operator and the Kolmogorov backward equation to derive optimal importance sampling and multilevel splitting estimators. By expressing an indicator function over a rare event in terms of the eigenfunctions of the stochastic Koopman operator, we directly approximate the associated zero-variance importance sampling estimator. We also devise efficient multilevel splitting schemes for SDEs by using the Koopman eigenfunctions to approximate the optimal importance function.

Stochastic dynamical systems can also be tools for solving problems in computational statistics. Creative uses of SDEs have been instrumental in developing efficient sampling methods for high-dimensional, non-Gaussian probability distributions. The second part of the thesis develops new sampling methods that employ judiciously constructed SDEs. We first present a framework for constructing *controlled* SDEs that can sample from a large class of probability distributions with Gaussian tails, in finite time. By choosing a linear SDE to be the uncontrolled reference system, we synthesize feedback controllers that drive the sampling of such distributions. We identify and approximate these controllers by solving only a static optimization problem.

Next, we develop novel approaches for accelerating the convergence of *Langevin dynamics*-based samplers. Reversible and irreversible perturbations of Langevin dynamics can improve the performance of Langevin samplers. We present the geometry-informed

irreversible perturbation (GiIrr) and show that it accelerates convergence of Riemannian manifold Langevin dynamics more than standard irreversible perturbations. We then propose the transport map unadjusted Langevin algorithm (TMULA), and show that the use of transport enables rapid convergence of the unadjusted Langevin algorithm for distributions that are not strongly log-concave. We also make connections between transport maps and Riemannian manifold Langevin dynamics to elucidate how transport maps accelerate convergence.

Thesis Supervisor: Youssef M. Marzouk
Title: Professor of Aeronautics and Astronautics

Thesis Committee Member: Tuhin Sahai
Title: Technical Fellow, Raytheon Technologies Research Center

Thesis Committee Member: Themistoklis Sapsis
Title: Associate Professor of Mechanical and Ocean Engineering

Thesis Committee Member: Konstantinos Spiliopoulos
Title: Associate Professor of Mathematics and Statistics, Boston University

Acknowledgments

Research is not an individual affair; the most valuable experiences of the last six years have been the connections and collaborations I have been able to cultivate and enjoy. There are many people to thank for making my graduate school experience fruitful and rewarding.

I am very grateful to my advisor, Professor Youssef Marzouk for giving me the opportunity to pursue graduate studies at MIT. He has given me the resources and freedom to dive into many interesting research directions on my own, and has enriched my graduate school experience by inviting interesting research visitors, and by encouraging me to attend conferences and workshops. I especially thank him for cultivating a research group which values collaboration, curiosity, and a supportive atmosphere. I am also deeply indebted to Dr. Tuhin Sahai who in many respects served as a co-advisor to me. I appreciate the time he has taken to listen and tolerate my research grumbles and for encouraging me to think outside the box for my research. I also thank him for being my internship supervisor when Raytheon was still UTRC.

I thank Professor Themistoklis Sapsis for his thoughtful feedback and suggestions on the thesis, and for our interesting discussions on rare events in ocean engineering and fluid mechanics. I am also thankful for Professor Konstantinos Spiliopoulos at Boston University for his wisdom and acumen on the technical aspects of rare events and stochastic analysis. In particular, I thank him for pointing me towards Langevin-based sampling methods which has led to many fruitful research ideas. I also thank him for being the external evaluator for my thesis proposal.

Thanks are also due to Professor Igor Mezić at UCSB and Professor Jonathan Weare at NYU for serving as external thesis readers and for providing valuable feedback that has enhanced the thesis.

I thank Dr. Quan Long for his collaboration on the multilevel splitting project. I am also very appreciative of my undergraduate research students, Joshua White and Karolina Podsada. Josh was instrumental in performing the numerical experiments related to studying the Ornstein-Uhlenbeck operator and multilevel splitting. Karolina

helped me better understand numerical methods for SDEs and SPDEs.

The friends I have made during my time at MIT have been indispensible. They have taught me so much about research, work, and life in general. This thesis would not have been possible without their technical advice, moral support, memes, and fun times outside of work. I thank Ricardo Baptista, Michael Brennan, Nisha Chandramoorthy, Yuankang Chen, Alex Feldstein, Chi Feng, Dan Fortunato, Mohammad Islam, Fengyi Li, Harriet Li, Friedrich Menhorn, Angxiu Ni, Elizabeth Qian, Tom R., Andrea Scarinci, Alessio Spantini, Sven Wang, Zheng Wang, and Shun Zhang.

I also thank Kate Nelson and Jean Sofronas for tirelessly supporting students in navigating through MIT's administrative curiosities.

Finally, I thank my parents Jie Wang and Yubao Zhang, and my brother Eric Zhu for their constant love and support.

This thesis was made possible by financial support from the DARPA EQUiPS program, the MathWorks Engineering Fellowship, and the ANSRE AFOSR MURI.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

This thesis celebrates the interplay between *computational statistics* and *computational dynamics*. Dynamical systems and their related computational methods are powerful tools for describing and studying phenomena that evolve in time. State-of-the-art methods for uncertainty quantification of dynamical systems employ tools from computational statistics and are crucial for predictive modeling and decision making. Concurrently, modern computational statistics methodology, such as Markov chain Monte Carlo, make use of dynamical systems to sample from complex, high-dimensional probability distributions. There are many areas of study that lie at the intersection of computational statistics and dynamics, including as data assimilation, stochastic optimal control, parameter inference for dynamical systems, nonlinear programming, model reduction, interacting particle systems, and many others. In this thesis, we focus on two topics: (1) rare event simulation for stochastic differential equations, and (2) sampling probability distributions with SDEs.

1.1.1 Rare event simulation for stochastic dynamical systems

Our first major theme centers on a framework for developing efficient sampling methods for computing the statistics of SDEs. In particular, we focus on estimating rare event

probabilities. Understanding and quantitatively characterizing rare phenomena is important to modeling, design, and decision making in a variety of science and engineering disciplines. Examples include studying the failure of materials [75], predicting the insolvency of financial institutions [46], understanding the occurrence of rogue waves [30, 36], estimating reaction rates in computational chemistry [92, 121], and assessing the reliability of aerospace systems [135]. Many of these examples involve dynamical systems forced by random noise, which is captured in the form of Brownian motion, and a key challenge is to compute the probabilities of noise-induced rare events and the predominant mechanisms by which they occur. These rare events are often associated with adverse outcomes and failures. Moreover, many complex engineered systems are often available as a black-box, developing a purely data-driven approach to quantify rare event probabilities and characterize their mechanisms is of interest for these challenging engineering system.

Rare event simulation for SDEs is particularly challenging for two reasons. It first requires a faithful model, i.e., one that exhibits the rare phenomena of interest with sufficiently accurate probability. Second, a computationally efficient methodology is needed to produce the rare event, i.e., to characterize the tails of the relevant distributions. Performing the latter also elucidates the pathways or *mechanisms* leading to a rare event.

For SDEs, expectations with respect to the induced path-space probability measures are, generally, difficult to compute directly. Hence Monte Carlo methods are often used to estimate these expectations instead. Simple Monte Carlo methods, while robust, are inefficient for estimating expectations sensitive to rare events [122]. Since rare events by definition occur infrequently, the variance of a simple Monte Carlo estimator can be very large relative to the quantity of interest. Furthermore, for rare events in SDEs that obey a large deviations principle, simple Monte Carlo methods require an exponentially increasing number of samples to maintain a constant relative error as the noise factor in the Brownian motion decreases linearly [37, 122]. For these reasons, a vast body of literature has focused on devising sampling methods that improve on simple Monte Carlo for rare event simulation [17].

Importance sampling for SDEs constitutes a major class of Monte Carlo methods for simulating rare events. Here, one simulates an alternative dynamical system whose trajectories reach the rare event more often. Each of these samples is then re-weighted according to its importance relative to the original SDE's distribution. These weights are given by the celebrated Girsanov theorem [85]. Multi-level splitting and subset simulation comprise a different class of adaptive Monte Carlo methods for rare events in which, over a series of iterations, one creates more unbiased, independent trajectories that tend towards the rare event of interest. Splitting methods were first conceived in [64], with more computationally efficient methods proposed recently [22, 127]. Subset simulation, on the other hand, was originally proposed in the engineering reliability literature [5] and has been widely adopted and improved upon by the civil engineering community [88]. While subset simulation can be used to estimate rare event probabilities in dynamical systems, it is typically used to sample static models. Links between subset simulation and sequential Monte Carlo are described in [20], and multilevel splitting has been extended to static and non-Markovian models in [12]. Importance sampling for SDEs, while simple to implement and easily parallelizable, is intrusive: in the context of SDEs, it requires altering the drift term of the model, which may be impossible when the model is given as a black box. In contrast, multilevel splitting methods are applicable in black-box settings and often more stable than their importance sampling counterparts [17].

In the first part of this thesis, we propose a novel framework for rare event simulation that uses tools originating from dynamical systems theory and combines them with importance sampling (IS) and multilevel splitting (MS). Specifically, we show that the *stochastic Koopman eigenfunctions* (sKO) associated with a given SDE can be used to accurately and efficiently approximate zero-variance IS estimators, and optimal MS estimators. The approach leverages recent developments in Koopman operator approximation techniques and only assumes that the SDE is amenable to numerical Koopman analysis.

The last decade has witnessed considerable interest in operator-theoretic and data-driven computational approaches for analyzing and manipulating nonlinear dynamical

systems. The Koopman operator is a linear mapping on the space of observables of a given dynamical system [18, 68, 81, 83]. Its existence provides a *global* linearization of the dynamics and enables spectral analysis for nonlinear systems. Moreover, the discovery that data-driven methods for dynamical systems such as dynamic mode decomposition (DMD) [106] (originally conceived in the fluid mechanics community) can effectively approximate spectral objects of the Koopman operator [103] has led to their further development and widespread application.

In nonlinear settings where the stochastic Koopman eigenfunctions cannot be derived analytically, we use dynamic mode decomposition (DMD) methods to approximate them numerically. There is no zero-variance estimator for multilevel splitting. However, the optimal implementation of multilevel splitting, whose variance attains the theoretical lower bound, is related to the solution of the KBE. Therefore, we also study the use of approximate solutions of the KBE with sKO eigenfunctions for multilevel splitting.

1.1.2 Stochastic differential equations for sampling

The second part of this thesis explores the use of stochastic differential equations for sampling. A common problem in statistics and machine learning is that of computing expectations with respect to complex probability distributions. These distributions frequently arise as posterior distributions in Bayesian statistics. Bayesian inference is a powerful framework for fusing observational data with prior knowledge to learn model parameters. Computing expectations with respect to a target posterior distribution arising from the Bayesian framework is often a challenging problem when the target is highly non-Gaussian. To solve this problem, particularly in high dimensions, one frequently resorts to sampling methods. Estimating these quantities efficiently via Monte Carlo requires computationally efficient schemes for producing samples that approximate the distribution. We explore two separate ways for using SDEs to solving problems in Bayesian inference.

Our first approach re-interprets our framework for rare event simulation. We develop a framework for constructing a family of *controlled* stochastic dynamical

systems that can *exactly* sample from a class of probability distributions with Gaussian tails on $\mathbb{R}^d$. Recently the theory of controlled diffusion processes has been gaining attention in statistics and machine learning. Given a reference stochastic differential equation (SDE), an initial distribution, and a target distribution, one aims to find a feedback control such that the marginal at some finite future time T is equal to the target distribution. Finding the optimal feedback control enables *exact sampling* of the target. This problem is also known as the Schrödinger bridge problem [107]. The optimal control is known by different names in different communities, including the *Doob h-transform* [38, 122] and the Föllmer drift [48, 120].

We present a special case of the approach when the reference system is a linear SDE, and show that the optimal control can be expressed in terms of the eigenfunctions of the reference system's Markov generator. Relating these eigenfunctions to the target distribution only requires solving a static optimization problem, instead of a dynamic problem in the optimal control formulation. The resulting controlled (nonlinear) SDE can produce samples from a broad class of target distributions, in parallel, and can be used for importance sampling, or for approximate inference.

A different, but well established, class of sampling methods using dynamical systems is based on the Langevin dynamics (LD), which uses the gradient of the log-target density to specify a SDE whose invariant distribution is the target distribution of interest. Long term averages over a single trajectory of the SDE can be then used to estimate expectations of interest by appealing to the ergodicity of the stochastic process. It is also known that certain perturbations to the LD can accelerate convergence of the dynamics to the stationary distribution. In [95] the authors show that suitable reversible and irreversible perturbations to diffusion processes can decrease the spectral gap of the generator and decrease the asymptotic variance of the estimators. One widely celebrated choice of *reversible* perturbation is the Riemannian manifold Langevin dynamics (RMLD) [53], in which one defines a Riemannian metric to alter the way distances and gradients are computed. The use of *irreversible* perturbations to accelerate convergence has also been well studied in a variety of contexts and general settings [61, 95, 96, 97]. By augmenting the drift of LD with a vector field that is

orthogonal to the LD drift will leave the invariant density unchanged while accelerating its convergence.

We propose a *state-dependent irreversible* perturbation of Riemannian manifold Langevin dynamics that is informed by the *geometry* of the manifold. This departs from existing literature, as the vector field of the resulting perturbation is *not* orthogonal to the original drift term. We observe that this geometry-informed irreversible perturbation accelerates convergence more than standard irreversible perturbations and, if desired, can be used in combination with the SGLD algorithm to exploit the computational savings of a stochastic gradient. Lastly, we explore how the transport maps-based Langevin samplers accelerate convergence through the perspective of reversible perturbations and through recent results on the convergence of unadjusted Langevin algorithms.

1.2 Thesis contributions and outline

We present our contributions in the following chapters. The first part of the thesis contains two chapters that establish a framework for efficient sampling for estimating expectations of interest in SDEs based on the Koopman operator and its associated numerical methods.

- Chapter 2 develops a way to construct efficient importance sampling estimators for rare event simulation in nonlinear SDEs by using the Koopman operator. Specifically, we propose using eigenfunctions of the stochastic Koopman operator to approximate the Doob transform, which is a feedback controller, for an observable of interest (e.g., associated with a rare event) which in turn yields an approximation of the corresponding zero-variance importance sampling estimator. In nonlinear settings where the stochastic Koopman eigenfunctions cannot be derived analytically, we use dynamic mode decomposition (DMD) methods to approximate them numerically. Numerical experiments demonstrate that even coarse approximations of a few eigenfunctions, where the latter are built from non-rare trajectories, can produce effective importance sampling schemes for

rare events.

- Chapter 3 extends the framework described in Chapter 2 to more robust rare event sampling schemes, specifically to multilevel splitting. In contrast to importance sampling, multilevel splitting is a black-box approach to rare event simulation that only requires sample trajectories of the model. Trajectories are incentivized to move towards the rare event where favorable trajectories are allowed to split. This approach requires a judicious partition of the state space defined by level sets of an importance function (sometimes referred to as the score function). We use the stochastic Koopman operator (sKO) and its related numerical methods to approximate the optimal importance function for splitting.

In the second part of the thesis, we present our contributions to efficient sampling methods for probability distributions using stochastic differential equations. We first study the construction of a class of *controlled* SDEs for sampling before considering novel reversible and irreversible perturbations that accelerate the convergence of Langevin dynamics-based sampling methods.

- In Chapter 4, we re-interpret the framework developed in Chapter 2 to create an approach for constructing controlled stochastic differential equations that exactly sample from a class of probability distributions with Gaussian tails. Given a target distribution and a reference SDE, the Doob h-transform produces a controlled stochastic process whose marginal at a finite time T will be equal to the target distribution. Our method constructs a reference linear SDE and uses the eigenfunctions of its associated Ornstein-Uhlenbeck (OU) operator to approximate the Doob h-transform. The control is approximated by projecting the ratio between the target density and the reference system's time T marginal onto the span of a finite set of OU eigenfunctions. This projection is performed by minimizing the Kullback-Leibler (KL) divergence from the marginal produced by the approximate control to the true target distribution. We relate our work to the Schrödinger bridge problem.

- In Chapter 5, we introduce a novel geometry-informed irreversible perturbation that accelerates convergence of the Langevin algorithm for Bayesian computation. Langevin dynamics (LD) are widely used in sampling high-dimensional, non-Gaussian distributions whose densities are known up to a normalizing constant. In particular, there is recent interest in the unadjusted Langevin algorithm (ULA) in which a single realization of LD is used to estimate expectations with respect to the target distribution. There exist perturbations to the Langevin dynamics that preserve its invariant measure while accelerating its convergence. Irreversible perturbations and reversible perturbations (such as Riemannian manifold Langevin dynamics (RMLD)) have separately been shown to improve the performance of Langevin samplers. We consider these two perturbations simultaneously by presenting a novel form of irreversible perturbation for RMLD that is informed by the underlying geometry. We demonstrate our approach on posterior distributions arising from Bayesian logistic regression and independent component analysis.

- In Chapter 6, we continue exploring perturbations for improving Langevin-based samplers. When the target distribution is not strongly log-concave, ULA is known to exhibit slow convergence. The use of transport maps have been shown to accelerate the Metropolis-adjusted Langevin algorithm (MALA) where the map creates non-Gaussian proposals and has been empirically shown to accelerate convergence to the target distribution. We show that, under certain conditions on the transport map and the target density, when a map is used in conjunction with ULA, we can obtain geometric convergence of the output process in the 2–Wasserstein distance even when the target distribution does not satisfy the typical conditions for such rapid convergence. Moreover, we also show that in the continuous-time setting, when a transport map is applied to LD, the result is a RMLD with a metric that is defined by the transport map. Finally, we make some connections between this approach and variational formulations of Bayesian inference.

In Appendix A, we study computing the eigenfunctions of the Ornstein–Uhlenbeck (OU) operator, which is the stochastic Koopman operator for linear SDEs. The computation of OU eigenfunctions supports the work in Chapters 2, 3, and 4. In Appendix C, we also provide a brief note on methods for simulating stochastic partial differential equations, which are used in Chapter 2.

1.2.1 Preprints

- Chapter 2 is based on work in [137], which is currently under review at the *Journal of Computational Physics*.
- Chapter 5 is based on work in [136], which is currently under review at *Statistics and Computing*.
- Appendix A is based on work in [138].

Part I

Sampling methods *for* stochastic dynamical systems

Chapter 2

A Koopman framework for importance sampling and rare event simulation

2.1 Introduction

In this chapter, we present a general framework for constructing Monte Carlo estimators of rare event probabilities, and of other expectations associated with rare events, in nonlinear stochastic differential equations (SDEs). In particular, we focus on approximating the zero-variance importance sampling estimator using the stochastic Koopman operator. By expressing the indicator function over a rare event in terms of the system's stochastic Koopman eigenfunctions, we can approximate the Doob h-transform and obtain a lower variance estimator.

Intuitively, importance sampling for SDEs can be interpreted as a stochastic optimal control problem. We want to find the most probable path that leads to the system to end up in a rare event. There are many established approaches to importance sampling for SDEs. For example, importance sampling have been further enhanced by large deviations theory [37, 123]. These methods exploit the large deviations principle as an alternative mechanism of characterizing rare events in dynamical

systems, inform the implementation of splitting and importance sampling, and provide theoretical guarantees on estimator efficiency [42, 122]. These methods also appear in the literature as so-called *instanton*-based sampling methods, where minimizers of the system's large deviations rate function describe how to push the system towards the rare event of interest [45, 79]. These enhancements are also related to *variational* approaches to importance sampling, where the alternative SDE is posed as the solution to a stochastic optimal control problem—which, in principle, can yield zero-variance estimators [57, 56, 66, 139]. The drawbacks of these approaches are also well-noted. Large deviations-based approaches for sampling are optimal in an asymptotic sense, but counter-examples have been constructed to show that they can lead to larger variance than applying direct Monte Carlo [54]. And the computational effort required to solve stochastic optimal control problems can be untenable in high-dimensional settings.

In this Chapter, we use the relationship between zero-variance sampling and the Koopman operator to show that DMD methods can be integrated with importance sampling to create new rare event simulation techniques. Our framework provides a systematic approach with *general* applicability. For example, existing rare event simulation techniques are often demonstrated on gradient systems, or on systems with normal dynamics. Our approach is also applicable to non-gradient systems, non-normal systems that display transient growth, and oscillatory dynamics. This flexibility is critical for extending efficient dynamic rare event simulation to realistic engineering problems [135].

A key feature of our approach is that we leverage the data-driven nature of Koopman numerics to provide insight into rare events via simulation of *non-rare trajectories*. The ability to resolve Koopman eigenfunctions near the rare event using non-rare trajectories enables computation of a biasing that "pushes" importance sampling trajectories into the rare event regions. We show that even *coarse approximations* of the Koopman eigenfunctions using non-rare trajectories can produce good importance sampling estimators for rare event simulation. While the training data do not need trajectories that exhibit the exact rare trajectory we are trying to simulate, it is

necessary to have training data that sufficiently covers the transition pathways to the rare event. In theory, these trajectories can be long or short, but we have found that long trajectories produce better approximation of the eigenfunctions. The method is asymptotically exact in the sense that as one employs a larger number of Koopman eigenfunctions, the variance of the corresponding importance sampling estimator tends towards zero. We provide a non-asymptotic analysis that describes how, under certain conditions, the second moment of the importance sampling estimator is bounded by a term that depends on how well the Koopman eigenfunctions approximate the observable of interest.

2.1.1 Problem setting and notation

Let $\{X_t\}_{t\in[0,T]}$ be a time-homogeneous diffusion process evolving according to the SDE,

$$\begin{cases} \mathrm{d}X_t &= \mathbf{A}(X_t)\,\mathrm{d}t + \mathbf{B}(X_t)\,\mathrm{d}W_t \\ X_0 &= x, \end{cases} \tag{2.1}$$

where X_t is an element of $\mathbb{R}^d$, $\mathbf{A}$ is a function from $\mathbb{R}^d$ to itself, $\mathbf{B}$ is a function from $\mathbb{R}^d$ to the space of $d \times r$ real-valued matrices, and W_t is a standard r-dimensional Brownian motion. To guarantee existence and uniqueness of a strong solution to the SDE, we assume the drift vector and diffusion matrix are locally Lipschitz in space [65]. We wish to estimate

$$\rho = \mathbb{E}[f(X_T)|X_0 = x] = \int_{\mathbb{R}^d} f(x)\pi_T(x)\,\mathrm{d}x, \tag{2.2}$$

where $f(x)$ is non-negative, and $\pi_T(x)$ is the probability density of the state at time T. Note that if $f(x)$ were an *indicator function over some rare event of interest* E, then ρ would equal the probability of the state being in region E at time T. We also assume that the system has an invariant distribution η_∞.

In the next section, we review some theoretical tools for importance sampling in

stochastic differential equations. In Section 2.3, we discuss the Koopman operator and related numerical methods, and we present our framework for constructing importance sampling estimators. In Section 2.4, we demonstrate the methodology on a range of illustrative stochastic dynamical systems. We analyze the variance of the importance sampling estimators produced by our methodology in Section 2.5. We discuss future work in Section 2.6.

2.2 Rare event simulation for SDEs

We start with an overview of analytical tools for studying stochastic differential equations, including the infinitesimal generator and the Kolmogorov equations. Much of this discussion is based on Karatzas and Shreve [65], Øksendal [85], and Pavliotis [90]. We also review importance sampling in the context of SDEs and describe related approaches to rare event simulation based on stochastic optimal control and large deviations theory.

2.2.1 Kolmogorov equations

Let X_t be defined by the SDE in (2.1). One of the primary tools for studying stochastic processes is the infinitesimal generator defined as

$$\mathcal{A}f = \lim_{t \to 0} \frac{\mathbb{E}[f(X_t)|X_0 = x] - f(x)}{t}, \tag{2.3}$$

for $f \in \mathcal{D}_{\mathcal{A}}$, where $\mathcal{D}_{\mathcal{A}}$ is the set of functions for which the above limit exists for all $x \in \mathbb{R}^d$. For SDEs, a closed form expression of the limit involves the drift and diffusion terms as follows,

$$\begin{aligned} \mathcal{A}f &= \langle \mathbf{A}(x), \nabla f \rangle + \operatorname{Tr}\big[\mathbf{Q}(x)\nabla^2 f\big], \\ &= \sum_{i=1}^{d} A_i(x) \frac{\partial f}{\partial x_i} + \sum_{i=1}^{d}\sum_{j=1}^{d} Q_{ij}(x) \frac{\partial^2 f}{\partial x_i \partial x_j}, \end{aligned} \tag{2.4}$$

where $\mathbf{Q}(x) = \frac{1}{2}\mathbf{B}(x)\mathbf{B}(x)^*$ and ψ is a twice-continuously differentiable function on $\mathbb{R}^d$. The infinitesimal generator appears in the Kolmogorov equations, which are two PDEs that describe the evolution of densities and statistics of a given SDE. The *Kolmogorov backward equation* (KBE) describes the time-evolution of expectations of functions of the state. Let $\Phi(t,x) = \mathbb{E}^{t,x}[f(X_T)] \coloneqq \mathbb{E}[f(X_T)|X_t = x]$ be defined on $t \in [0,T]$, where $T > 0$. Then

$$\begin{cases} \dfrac{\partial \Phi}{\partial t} + \mathcal{A}\Phi = 0 \\ \Phi(T,x) = f(x). \end{cases} \tag{2.5}$$

The *Kolmogorov forward equation* (KFE), also known as the Fokker–Planck equation, describes the evolution of the probability density function of the state. The equation is found by considering the L^2-adjoint of the infinitesimal generator. Let $\pi(t,x)$ be the probability density of X_t. Then

$$\begin{cases} \dfrac{\partial \pi}{\partial t} = \mathcal{A}^*\pi(t,x) \\ \pi(0,x) = \pi_0(x) \end{cases} \tag{2.6}$$

where the adjoint is

$$\mathcal{A}^*\pi = -\nabla \cdot (\mathbf{A}(x)\pi) + \mathrm{Tr}\big[\nabla^2(\mathbf{Q}(x)\pi)\big]. \tag{2.7}$$

Theoretically, expectations such as (2.2) can be found via a direct solution of the KBE. The quantity of interest is simply an evaluation of the solution: $\rho = \Phi(0,x)$. However, solving the KBE exactly is expensive and increasingly intractable as the dimension of the state space grows. Furthermore, when one is interested in quantities such as rare event probabilities, the required solution accuracy typically becomes prohibitive. For this reason, we turn to sampling methods, in which multiple independent simulations of an SDE are performed to estimate expectations through a sample average. While a direct solution of the Kolmogorov equations may not be feasible, in what follows, we show that these equations can be used to approximate zero-variance estimators.

2.2.2 Importance sampling for SDEs

We now review some basic notions of Monte Carlo and importance sampling methods for SDEs. Let $\mathbb{P}$ be the path-space measure induced by the SDE in (2.1). A simple Monte Carlo method for estimating ρ involves generating M independent simulations of the SDE, evaluating the function of interest $f(x)$ at the end of each sample path, and computing the sample average. We then have

$$\rho \approx \hat{\rho} = \frac{1}{M} \sum_{i=1}^{M} f(X_T^{(i)}), \tag{2.8}$$

where the samples $X^{(i)}$ are drawn independently from $\mathbb{P}$. The efficiency of a Monte Carlo estimator is typically evaluated by considering its variance and relative error (also known as the coefficient of variation, i.e., the standard error divided by the quantity of interest) [4]. They are, respectively,

$$\mathbb{V}[\hat{\rho}] = \frac{1}{M} \operatorname{Var}[f(X_T)], \tag{2.9}$$

$$r_e = \frac{1}{\rho} \sqrt{\mathbb{V}[\hat{\rho}]}. \tag{2.10}$$

A good unbiased estimator should have low variance, but when estimating a small ρ such as a rare event probability, relative error is the better metric. This is because r_e can still be large if ρ is orders of magnitude smaller than $\operatorname{Var}[\hat{\rho}]$. In other words, our goal is to ensure that the standard deviation of the estimator scales in proportion with the probability of interest. The relative error per sample, $r_e\sqrt{M}$, is a useful standardized measure of performance as it is independent of the sample size M [104].

The inefficiency of simple Monte Carlo methods is clear when used to estimate rare event probabilities. Let $f(x) = \mathbb{1}_E(x)$ where $E \subset \mathbb{R}^d$ is a region of phase (or observable) space visited infrequently. The variance and relative error of the estimator

are,

$$\mathrm{Var}[\hat{\rho}] = \frac{\rho - \rho^2}{M} \approx \frac{\rho}{M},$$
$$r_e \approx \frac{1}{\sqrt{M\rho}}.$$

We can see that the number of samples required to keep the relative error below 1 is $O(1/\rho)$. This task is particularly intractable when it is computationally expensive to procure samples from the dynamical system. In simple Monte Carlo, the only way one can reduce the variance of the estimator is by increasing the number of samples in each estimate. Variance reduction methods pursue different mechanisms for reducing the variance beyond simply increasing the number of samples.

One common variance reduction approach is *importance sampling*, which involves drawing samples from an alternative probability measure, $\mathbb{Q}$, that is absolutely continuous with respect the original probability distribution, such that the variance of the resulting estimator is reduced. To account for the bias introduced when sampling from the alternative probability distribution, each sample is weighted according its relative importance with respect to the original measure $\mathbb{P}$. In particular,

$$\hat{\rho}_{IS} = \frac{1}{M}\sum_{i=1}^{M} f(\tilde{X}_T^{(i)})\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(\tilde{X}^{(i)}), \tag{2.11}$$

where $\tilde{X}^{(i)}$ are drawn independently from $\mathbb{Q}$. The variance of this estimator is dependent on the product of the function $f(x)$ and the likelihood ratio between $\mathbb{P}$ and $\mathbb{Q}$, and on the number of samples drawn:

$$\mathrm{Var}[\hat{\rho}_{\mathrm{IS}}] = \frac{1}{M}\,\mathrm{Var}_{\mathbb{Q}}\left[f(\tilde{X}_T)\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right]. \tag{2.12}$$

Thus, designing a measure $\mathbb{Q}$ provides an additional mechanism to reduce the variance of the sampling method. For SDE systems, the only admissible class of $\mathbb{Q}$ is induced by another SDE system $\{\tilde{X}\}_{t\in[0,T]}$ with the same diffusion term as the original SDE

and a different drift term [85, 112]:

$$\begin{cases} \mathrm{d}\tilde{X}_t = \Big[\mathbf{A}(\tilde{X}_t) + \mathbf{B}(\tilde{X}_t)u(t,\tilde{X}_t)\Big]\mathrm{d}t + \mathbf{B}(\tilde{X}_t)\mathrm{d}W_t \\ \tilde{X}_0 = x. \end{cases} \tag{2.13}$$

Here $u(t,x)$ is called the biasing function. This function serves as a feedback controller that guides the system such that the resulting importance sampling estimator has lower variance. The likelihood ratio is now given by Girsanov's theorem [85, 65]:

$$Z(\tilde{X}) \equiv \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(\tilde{X}) = \exp\bigg(-\int_0^T \langle u(t,\tilde{X}_t), \mathrm{d}W_t\rangle - \frac{1}{2}\int_0^T \|u(t,\tilde{X}_t)\|^2 \mathrm{d}t\bigg). \tag{2.14}$$

The task now is to choose $u(t,x)$ such that the variance of the resulting importance sampling estimator is smaller, or better yet, zero. Assuming that $f(x)$ is twice-continuously differentiable and strictly positive, there exists a choice of $u(t,x)$ that leads to a *zero-variance* importance sampling estimator. This choice is the celebrated *Doob h-transform* [101, 105, 122].

Theorem 1 (Doob h-transform[1])**.** *Let $f \in \mathcal{C}^2$ be strictly positive. Let $\Phi(t,x) = \mathbb{E}^{t,x}[f(X_T)]$ be the solution to*

$$\begin{cases} \dfrac{\partial \Phi}{\partial t} + \mathcal{A}\Phi = 0, \\ \Phi(T,x) = f(x). \end{cases} \tag{2.15}$$

Then using the biasing function

$$u(t,x) = \boldsymbol{B}^*(x)\nabla \log \Phi(t,x) \tag{2.16}$$

[1]As written, Theorem 1 does not apply when f is an indicator function. This is an artifact of the simple way we have chosen to express the result. It is straightforward to modify it by conditioning on the event that X_T enters a particular region; indeed, the Doob transform was originally derived just for conditioned processes [101]. Also, our numerical experiments will mollify and positivize the indicator function in constructing numerical approximations of the Doob transform, such that Theorem 1 applies directly.

in (2.13) *will satisfy*

$$f(\tilde{X}_T)\exp\left[-\int_0^T \langle u(t,\tilde{X}_t), dW_t\rangle - \frac{1}{2}\int_0^T \|u(t,\tilde{X}_t)\|^2 dt\right] = \Phi(0,x). \tag{2.17}$$

While the Doob h-transform is a standard result [101, 105], we provide a proof of the particular form presented in Theorem 1 to elucidate how it is important in the construction of our importance sampling estimator.

Proof. We compute the stochastic integral

$$\int_0^T \langle u(t,\tilde{X}_t), \mathrm{d}W_t\rangle.$$

Let $g(t,x) = \log\Phi(t,x)$, and apply Itô's formula:

$$\begin{aligned}
\mathrm{d}g =& \frac{\partial}{\partial t}\log\Phi(t,\tilde{X}_t)\mathrm{d}t + \langle\nabla\log\Phi(t,\tilde{X}_t), \mathrm{d}\tilde{X}_t\rangle + \frac{1}{2}\mathrm{Tr}\Big[\nabla^2[\log\Phi(t,x)](\mathrm{d}\tilde{X}_t)(\mathrm{d}\tilde{X}_t)^*\Big] \\
=& \frac{1}{\Phi}\frac{\partial}{\partial t}\Phi\,\mathrm{d}t + \left\langle\frac{1}{\Phi}\nabla\Phi, \mathbf{A}(\tilde{X}_t) + \mathbf{B}\mathbf{B}^*\frac{\nabla\Phi}{\Phi}\right\rangle\mathrm{d}t + \left\langle\frac{\nabla\Phi}{\Phi}, \mathbf{B}\,\mathrm{d}W_t\right\rangle \\
&+ \frac{1}{2}\mathrm{Tr}\left[\mathbf{B}\mathbf{B}^*\frac{\nabla^2\Phi}{\Phi}\right]\mathrm{d}t - \frac{1}{2}\mathrm{Tr}\left[\mathbf{B}\mathbf{B}^*\frac{(\nabla\Phi)(\nabla\Phi)^*}{\Phi^2}\right]\mathrm{d}t \\
=& \frac{1}{\Phi}\left(\frac{\partial}{\partial t}\Phi + \langle\nabla\Phi, \mathbf{A}(\tilde{X}_t)\rangle + \frac{1}{2}\mathrm{Tr}\big[\mathbf{B}\mathbf{B}^*\nabla^2\Phi\big]\right)\mathrm{d}t + \frac{1}{2}\left\langle\frac{\mathbf{B}^*\nabla\Phi}{\Phi}, \frac{\mathbf{B}^*\nabla\Phi}{\Phi}\right\rangle\mathrm{d}t \\
&+ \left\langle\frac{\nabla\Phi}{\Phi}, \mathbf{B}\,\mathrm{d}W_t\right\rangle \\
=& \frac{1}{\Phi}\left(\frac{\partial\Phi}{\partial t} + \mathcal{A}\Phi\right)\mathrm{d}t + \frac{1}{2}\|\mathbf{B}^*\nabla\log\Phi(t,\tilde{X}_t)\|^2\,\mathrm{d}t + \langle\mathbf{B}^*\nabla\log\Phi(t,\tilde{X}_t), \mathrm{d}W_t\rangle \\
=& \frac{1}{2}\|u(t,\tilde{X}_t)\|^2\mathrm{d}t + \langle u(t,\tilde{X}_t), \mathrm{d}W_t\rangle.
\end{aligned}$$

This implies that

$$\int_0^T \langle u(t,\tilde{X}_t), dW_t\rangle = \log\Phi(T,x) - \log\Phi(0,x) - \frac{1}{2}\int_0^T \|u(t,\tilde{X}_t)\|^2\mathrm{d}t.$$

Plugging this into (2.17), we have

$$\begin{aligned} f(\tilde{X}_T) \exp\Big[- \int_0^T \langle u(t, \tilde{X}_t), \mathrm{d}W_t \rangle - \frac{1}{2} \int_0^T \| u(s, \tilde{X}_s) \|^2 \mathrm{d}s \Big] \\ &= f(\tilde{X}_T) \exp\Big[\log \Phi(0, x) - \log \Phi(T, \tilde{X}_T) \Big] \\ &= f(\tilde{X}_T) \frac{\Phi(0, x)}{f(\tilde{X}_T)} \\ &= \Phi(0, x). \end{aligned}$$

Thus, the biasing leads to a zero variance estimator. □

For more background on the Doob h-transform, we refer the reader to, e.g., [101, 105]. This choice of biasing results in a zero-variance estimator for ρ since $\rho = \Phi(0, x)$. This result should not be surprising: having access to the exact solution to the KBE enables construction of a Monte Carlo estimator with zero variance, since an evaluation of the solution is, itself, a zero-variance estimator. Though this relationship might seem tautological, it provides useful insights for devising efficient rare event simulation techniques.

Previous approaches recast this problem in terms of optimal control. By defining a new function $U(t, x) = -\log \Phi(t, x)$, one can obtain a PDE for $U(t, x)$ by performing a change of variables on the KBE. The resulting PDE is known as a *stochastic Hamilton–Jacobi–Bellman* (HJB) equation, which can be reformulated as a stochastic optimal control problem. In [57], the authors opt to solve the stochastic optimal control problem directly by using this formulation in conjunction with the Donsker–Varadhan variational formula. One can further recast the problem in terms of the solution of a system of forward-backward SDEs [66]. This approach also admits a cross-entropy interpretation for importance sampling for SDEs [139].

A similar approach incorporates the theory of large deviations, specifically the Freidlin–Wentzell theory for small noise diffusions [49]. Here, one considers a noise parameter ϵ that scales the diffusion term, by replacing $\mathbf{B}(x)$ with $\sqrt{\epsilon}\mathbf{B}(x)$. Then by considering the variable transformation $U^\epsilon(t, x) = -\epsilon \log \Phi(t, x)$ and sending ϵ to zero, one obtains a Hamilton–Jacobi equation whose solution is related to the

large deviations rate function of the system [122]. It was found that *subsolutions* of this Hamilton–Jacobi equation [41, 42, 43], for diffusion processes on $\mathbb{R}^d$ and in function space [104], result in provably asymptotically efficient estimators. However, the drawback of this approach is that the subsolution of the Hamilton-Jacobi equation must be "guessed," which is not always straightforward. Note that exact solutions of the resulting deterministic optimal control problem lead to strongly efficient estimators for SDEs [122].

Our approach, described below in Section 2.3, will avoid both of the above reformulations by directly computing *approximate Doob transforms* using approximate solutions to the KBE. These solutions of the KBE will be expressed in terms of the eigenfunctions of the stochastic Koopman operator. Our approach can also be related to the work of [27], in which the authors combine trajectory data from molecular dynamics simulations with nonlinear manifold learning techniques to inform the exploration of rare regions of state space. However, their technique is restricted to gradient systems for computational chemistry applications.

2.2.3 Related rare event problems

The problem posed in (2.2) is just one of many scenarios that are of interest in rare event simulation. In this chapter, we only consider the problem of the state being in some region of interest at some fixed future time T. Another common problem is to compute the probability of entering some region, E, before another, F: hence $\mathbb{P}(X_\tau \in E)$, where $\tau = \inf\{t > 0 : X_t \in E \cup F\}$. A variation of this problem considers path-dependent quantities, which involve functionals of sample trajectories. These problems are well-studied in the computational chemistry community, where one seeks rare paths between long-lived molecular configurations [56, 92, 121]. This quantity of interest is associated with the solution of a boundary value problem, and its approximation can also be used for sampling. The application of data-driven dynamical systems methodologies to this problem has been studied in [115].

Another quantity of interest is the probability of entering some set of interest within a *fixed* finite time interval, i.e., $\mathbb{P}(\tau \leq T)$ where $\tau = \inf\{t > 0 : X_t \in E\}$. This

problem is associated with escaping from attracting sets of a dynamical system and is well-studied in [41, 110]. In this case, asymptotically efficient importance sampling estimators are designed by considering an initial-boundary value problem associated with the KBE.

2.3 Importance sampling using the Koopman operator

We first review the deterministic and stochastic Koopman operators, and discuss how they can be used to approximate expectations and probabilities. We then describe how we will use the stochastic Koopman operator to construct importance sampling schemes for SDEs.

2.3.1 The Koopman operator and its generator

A traditional approach to analyzing dynamical systems involves simulating the evolution of *states*. The Koopman operator [68] provides an alternative perspective: it represents the dynamical system in terms of the evolution of *observables*. The key advantage is that the evolution of observables is linear even when the underlying system is nonlinear, thus enabling spectral analysis of nonlinear systems [18].

Let x_t be an autonomous dynamical system on $\mathbb{R}^d$ evolving according to $\dot{x} = a(x)$. Let F^t be the flow map; that is, if x_0 is the initial condition, then $x_t = F^t x_0$. Let $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ be an observable in some space of functions $\mathcal{H}$. The Koopman operator (KO) is defined as

$$\mathcal{K}^t f(x) = (f \circ F^t)(x). \tag{2.18}$$

It is trivial to show that the KO is linear even when the dynamical system is nonlinear. This property allows one to study the eigenfunctions and eigenvalues of the operator. A function $\phi(x)$ is a Koopman eigenfunction if it satisfies $\mathcal{K}^t\phi(x) = e^{\lambda t}\phi(x)$, where

λ is the corresponding Koopman eigenvalue. In stochastic calculus, the stochastic Koopman operator is known as the Markov semigroup operator of the SDE [90].

The stochastic Koopman operator (sKO) is defined in a similar fashion [31]. We focus our attention on random dynamical systems that evolve according to SDEs as defined in (2.1). Let $\{X_t\}_{t\in[0,T]}$ be a stochastic process and f be a twice continuously differentiable real-valued observable, respectively. Then the stochastic Koopman operator is defined as,

$$\mathcal{K}^t f(x) = \mathbb{E}[f(X_t)|x_0 = x] = \mathbb{E}^{0,x}[f(X_t)], \tag{2.19}$$

where the expectation is taken over the distribution of the state of the stochastic process at time t. Analogous to the deterministic setting, the sKO is also linear, leading to the spectral analysis of nonlinear SDEs. The evolution of the expectation of the sKO's eigenfunctions at future times is simple to determine. If $\phi(x)$ is an eigenfunction of the sKO, then $\mathbb{E}[\phi(X_t)|X_0 = x] = e^{\lambda t}\phi(x)$. Thus, the time evolution of certain observables of the dynamical system can be determined computationally.

2.3.2 Approximating expectations and probabilities

Assuming that the sKO eigenfunctions exist and form a basis for a suitable function space, expectations and probabilities associated with an SDE can, in principle, be calculated from all the eigenfunctions. Specifically, we can write the expectation of an observable at some fixed time in terms of the expectations of the sKO eigenfunctions, by first expressing the observable as a *linear combination* of these eigenfunctions.[2]

A finite collection of eigenfunctions can thus provide an approximation to the expectations and probabilities of interest. Let f represent some observable of interest

[2]For SDEs that admit an invariant measure and whose generators are compact and self-adjoint, the spectral theorem guarantees the existence of eigenvalues and a complete orthonormal set in $L^2(\eta_\infty)$, where η_∞ is the invariant measure. A frequently studied class of systems that admits a complete set of eigenfunctions are reversible diffusions. One example of a reversible diffusion occurs when the drift term is the gradient of a potential function and the diffusion matrix is the identity. In these cases, the solutions to the Kolmogorov equations can be found via eigenfunction expansions. See [90] for further details. Irreversible OU processes with invariant measure ν have also been shown to admit a complete basis of eigenfunctions on $L^p(\nu)$ for $p > 1$ [82].

and $\{\phi_i(x)\}_{i=1}^N$ be a collection of N eigenfunctions of the sKO with corresponding eigenvalues $\{\lambda_i\}_{i=1}^N$. Approximating the observable in terms of the eigenfunctions gives

$$f(x) \approx \sum_{i=1}^{N} f_i \phi_i(x), \tag{2.20}$$

and hence,

$$\begin{aligned} \mathbb{E}[f(X_t)|X_0 = x] &\approx \sum_{i=1}^{N} f_i \, \mathbb{E}^{0,x}[\phi_i(X_t)], \\ &= \sum_{i=1}^{N} f_i \mathcal{K}^t \phi_i(x) \\ &= \sum_{i=1}^{N} f_i e^{\lambda_i t} \phi_i(x). \end{aligned} \tag{2.21}$$

For rare event probabilities, it suffices to replace f with an indicator function over the rare set of interest; that is, to compute $\mathbb{P}(X_T \in E|X_0 = x)$, one would choose $f(x) = \mathbb{1}_E(x)$. In general, making this approximation accurate may require *accurately* computing *many* sKO eigenfunctions, which may not be practical in most settings. Instead we can combine this idea with importance sampling, as follows.

The sKO eigenfunctions can be used to create approximate solutions to the Kolmogorov backward equation. For continuous-time autonomous dynamical systems, the set of stochastic Koopman operators $\{\mathcal{K}^t\}_{t\in[0,\infty)}$ form a one parameter semigroup indexed by time. All elements of the semigroup share the same eigenfunctions, with varying eigenvalues depending on their parameter value. The generator of the semigroup is *identically* the infinitesimal generator of the SDE. That is, the generator of the sKO semigroup is exactly the evolution operator of the KBE. While this connection has been studied in stochastic analysis since the time of Kolmogorov, this connection is made most explicit in [31].

With this knowledge, we can construct importance sampling estimators for nonlinear SDEs. Observe that (2.21) provides an approximation to the quantity of interest in (2.2). Rather than using it directly to estimate the probability of the rare event,

we use it to approximate the Doob transform. Observe that

$$\tilde{\Phi}(t,x) = \sum_{i=1}^{N} f_i e^{\lambda_i (T-t)} \phi_i(x), \tag{2.22}$$

is an approximate solution to the KBE in (2.5). Then we can use the approximate Doob transform,

$$\tilde{u}(t,x) = \mathbf{B}(x)^* \frac{\sum_{i=1}^{N} f_i e^{\lambda_i (T-t)} \nabla \phi_i(x)}{\sum_{i=1}^{N} f_i e^{\lambda_i (T-t)} \phi_i(x)}, \tag{2.23}$$

to construct a new importance sampling scheme via (2.13) and (2.14). Intuitively, if $\tilde{\Phi}$ is a good approximation of the true solution, then the approximate Doob transform will be a good approximation to the true Doob transform, with the guarantee that if there exists a complete set of eigenfunctions, then the estimator will have zero variance as $N \to \infty$.

In practice, this framework offers considerable flexibility. While (2.21) provides an approximation to the quantity of interest, the errors introduced by truncation, and any additional errors resulting from numerical approximations of the eigenfunctions themselves, cannot easily be characterized. Instead, using the approximation within the Doob transform allows us to resolve these errors through Monte Carlo simulation. Our numerical experiments will demonstrate that even *crude* approximations of a *few* sKO eigenfunctions, where the latter are built from non-rare trajectories, can be used to build effective importance sampling methods for rare event probabilities. Moreover, the dynamics of the controlled SDE system (2.13) naturally reveal the most likely paths to the rare event.

Next we discuss numerical techniques for approximating the sKO eigenfunctions.

2.3.3 Dynamic mode decomposition methods

Dynamic mode decomposition (DMD) methods are a class of data-driven methods that can approximate eigenvalues and eigenfunctions of a (stochastic) dynamical system's (stochastic) Koopman operator. The original DMD method was presented in [106]

as means of model reduction for complex fluid flows. Low-dimensional behavior was extracted from time series data comprising snapshots of high-fidelity fluid dynamics simulations. The connection between DMD and the spectral objects of the Koopman operator was made clear by [103, 118], and there has since been considerable interest in developing more effective and efficient DMD methodologies and variants.

DMD methodologies typically use only sample trajectories of the system to approximate the Koopman eigenvalues and eigenfunctions, by indirectly approximating the infinitesimal generator [31, 133]. To avoid introducing errors due to these approximations of the generator, we use the analytical form of the SDE, which in turn provides access to the exact form of the generator of the sKO semigroup. In particular, we approximate Koopman eigenfunctions and eigenvalues using a recently developed variant of DMD called *infinitesimal generator extended dynamic mode decomposition* (gEDMD) [67]. The approach is based on using the stochastic Koopman generator in (2.3) directly. We summarize the main steps in the approach here.

Fix a set of test points $\{x_i\}_{i=1}^m$ drawn from a probability measure μ and a set of twice continuously differentiable basis functions $\{\psi_k(x)\}_{k=1}^n$.[3] Suppose the stochastic process $\{X_t\}$ evolves according to (2.1). The main idea is to project the action of the Koopman generator onto the basis functions. Following the notation of [67], let $\psi(x) = [\psi_1(x), \ldots, \psi_n(x)]^T$, define $\mathrm{d}\psi_k(x) \coloneqq (\mathcal{A}\psi_k)(x)$, and define

$$d\Psi_X = \begin{bmatrix} \mathrm{d}\psi_1(x_1) & \cdots & \mathrm{d}\psi_1(x_m) \\ \vdots & \ddots & \vdots \\ \mathrm{d}\psi_n(x_1) & \cdots & \mathrm{d}\psi_n(x_m) \end{bmatrix} \quad \Psi_X = \begin{bmatrix} \psi_1(x_1) & \cdots & \psi_1(x_m) \\ \vdots & \ddots & \vdots \\ \psi_n(x_1) & \cdots & \psi_n(x_m) \end{bmatrix}. \tag{2.24}$$

Let K be the finite dimensional representation of $\mathcal{A}$. The task is to find the matrix $K \in \mathbb{R}^{n\times n}$ such that the residual $\|\mathrm{d}\Psi_X - K\Psi_X\|_F$ is minimized, where $\|\cdot\|_F$ is the Frobenius norm. Each column of K is the solution to a least-squares problem, and it can be shown that $K = \mathrm{d}\Psi_X\Psi_X^+$, where $^+$ denotes the pseudoinverse. Furthermore, [67] shows that as the number of test points $m \to \infty$, this DMD method converges to a

[3] In our numerical experiments, we find that collecting test points from sample trajectories tends to produce better results (when validated on separate testing data) than prescribing some arbitrary measure μ.

Galerkin projection onto the span of the basis functions with respect to μ. Specifically, it is shown that,

$$K = \mathrm{d}\Psi_X \Psi_X^+ = (\mathrm{d}\Psi_X \Psi_X^T)(\Psi_X \Psi_X^T)^+ = \widehat{A}\widehat{G}^+, \tag{2.25}$$

where

$$\widehat{A} = \frac{1}{m}\sum_{i=1}^{m} \mathrm{d}\psi(x_i)\psi(x_i)^T, \quad \widehat{G} = \frac{1}{m}\sum_{i=1}^{m} \psi(x_i)\psi(x_i)^T. \tag{2.26}$$

And as the number of test points goes to infinity,

$$\lim_{m\to\infty} \widehat{A}_{ij} = \int (\mathcal{A}\psi_i)(x)\psi_j(x)\,\mathrm{d}\mu, \qquad \lim_{m\to\infty} \widehat{G}_{ij} = \int \psi_i(x)\psi_j(x)\,\mathrm{d}\mu. \tag{2.27}$$

The quality of the approximated eigenfunctions and eigenvalues will depend on the choice of basis functions and test point measure μ. We discuss our choices of basis functions and μ within the numerical examples of Section 2.4.3 for nonlinear stochastic systems; there, we also describe how we validate the resulting eigenfunction approximations. We summarize gEDMD in Algorithm 1.

Algorithm 1: Infinitesimal generator extended dynamic mode decomposition (gEDMD)

Input: SDE $dX_t = \mathbf{A}(X_t)\mathrm{d}t + \mathbf{B}(X_t)\mathrm{d}W_t$, Basis functions $\{\psi_j(x)\}_{j=1}^n$, measure μ
Output: Stochastic Koopman eigenfunctions $\{\phi_i(x)\}$ and eigenvalues $\{\lambda_i\}$

1: Obtain test points $\{x_i\}_{i=1}^m$ from measure μ
2: Evaluate $\{\psi_j(x)\}_{j=1}^n$ and $\{\mathcal{A}\psi_j(x)\}_{j=1}^n$ at the test points
3: Form matrices $\widehat{A}$, $\widehat{G}$, and K in (2.25) and (2.26)
4: Compute eigenvalues $\{\lambda_i\}_{i=1}^n$ and eigenvectors of $\{v_i\}_{i=1}^n$ of K
5: Eigenfunctions are $\phi_i(x) = v_i^T\psi(x)$.

Lastly, we discuss the potential for using other DMD methods. Our primary reason for selecting generator EDMD is that we have direct access to the generator, and would like to exploit it. Other extended dynamic mode decomposition (EDMD) methods could also be applicable [133]. A drawback of EDMD and its variants, however, is that they require a judiciously chosen basis, which can be difficult to devise in the

purely data-driven setting. There are methods such as (stochastic) Hankel DMD [3, 31] that do *not* require a dictionary of functions, and instead use delay embedding to generate a suitable dictionary. While such methods can accurately approximate the eigenfunctions at the test points, evaluating gradients of the eigenfunctions is more difficult. For our importance sampling approach, we need the ability to evaluate the gradient of the eigenfunctions *cheaply* and at a variety of *non-test* input points. Evaluating the gradients of Hankel DMD eigenfunction approximations is cumbersome, as it either requires interpolation or the solution of adjoint equations.

2.3.4 Approximating observables by eigenfunctions

Given a finite collection of sKO eigenfunctions, we can approximate solutions to the KBE, and hence the Doob transform, without having to solve the stochastic optimal control problems associated with existing rare event sampling methods. Computing these sKO eigenfunctions, as described in the previous section, is the first numerical challenge of our approach. The second challenge is to approximate the observable f as a linear combination of the eigenfunctions. We tackle this very simply, using linear regression with a least-squares objective. One may wonder if a projection of the indicator over the eigenfunctions would produce better results.

Formulating this regression problem precisely, and ensuring that the results can be used to define an appropriate biasing function via (2.23), requires resolving two issues. First is the choice of regression points. We construct the regression problem using the *same* point set used to approximate the sKO eigenfunctions, as described in the previous subsection. For rare event simulation, i.e., $f(x) = \mathbb{1}_E(x)$, properly representing the indicator function demands that some regression points lie inside the event of interest E. In our approach, we simulate many trajectories of the original system with different initial conditions throughout the domain and then subsample each trajectory to generate the regression (and EDMD) points. We assume that the user knows where the rare event E lies in state space, but has little idea *how* the system reaches it. To ensure that we have regression points inside E, we begin many of the sample trajectories inside the event of interest. In our numerical experiments,

for instance, the initial conditions are uniformly spaced over some subset of the state space that contains a portion of the rare event and the initial condition x. (Further details are given in Section 2.4.)

Now suppose that $\{\phi_i\}_{i=1}^N$ are the computed sKO eigenfunctions, and let $\{x_j\}_{j=1}^m$ denote the regression points. Let $\mathbf{f} = (f_1, \ldots, f_N) \in \mathbb{R}^N$ be the expansion coefficients in (2.20), $\mathbf{F} \in \mathbb{R}^m$ be evaluations of f at the regression points, and $\mathbf{C} \in \mathbb{R}^{m \times N}$ be the design matrix with $\mathbf{C}_{ji} = \phi_i(x_j)$. We then solve the least-squares problem,

$$\min_{\mathbf{f} \in \mathbb{R}^N} \|\mathbf{F} - \mathbf{C}\mathbf{f}\|_2^2 \,. \tag{2.28}$$

Next, recall that the Doob transform requires the approximate KBE solution (2.22) to be strictly positive. This property is not guaranteed by linear regression onto the eigenfunctions. One could add positivity constraints at the regression points to (2.28), but instead we correct "afterwards" by adding a constant to the approximate KBE solution produced by the regression. This correction does not impact the consistency of the sampling approach, because the constant function is always an sKO eigenfunction. The value of the approximated observable $\tilde{f}(x) = \sum_{i=1}^N f_i \phi_i(x)$ at each of the regression points can be found by computing $\mathbf{C}\mathbf{f}$. Assuming that the regression points sufficiently sample the relevant parts of the state space, we simply take the minimum of these values, denoted by $-\varepsilon$, and replace the coefficient f_1 of the constant sKO eigenfunction with $f_1 + \max(\varepsilon, 0)$.

It is important to note that adding a positive constant to $\tilde{f}$ will not affect the *direction* of the biasing function $\tilde{u}(t, x)$. The magnitude of $\tilde{u}$ will be diminished, however, since adding a constant increases the magnitude of the denominator in (2.23). This correction may thus cause the biasing to be too small to push the state into the rare event. To address this issue, we scale the biasing function by a multiplicative factor $c \geq 1$, to ensure that a sufficient fraction of trajectories reach the rare event when performing importance sampling; our final biasing is thus $\tilde{u}(t, x) = c\mathbf{B}(x)^* \nabla \log \tilde{\Phi}(t, x)$. In practice, we adjust c after finding the Doob transform, by simulating small batches of the controlled system with different c values and choosing a value such that a

sufficient fraction of samples (e.g., 0.2-0.6) reach the rare event. If c is chosen too small, not many samples will reach the rare event and the resulting estimator will have a large variance. If c is chosen too large, then too many samples will be pushed deeply into the rare event, but the resulting weights will be small. Since the estimator is unbiased, this implies that there will be at least one sample with a very large weight, which again implies that the estimator will have a large variance. We have observed that when the Doob transform is well approximated, the best factor c is close to one, meaning that essentially no multiplicative correction is needed. In Section 2.4.1, we observe the impact of c, and justify our procedure for choosing its value, through an example.

Our approach is summarized in Algorithm 2. In the next section, we provide further details on our implementation of the regression method and explore how the numerical choices above affect the performance of importance sampling.

Algorithm 2: Approximating the Doob transform.

Input: SDE $dX_t = \mathbf{A}(X_t)\mathrm{d}t + \mathbf{B}(X_t)\mathrm{d}W_t$ and observable $f(x)$
Output: Approximate Doob transform $\tilde{u}(t, x)$

1: Generate test points $\{x_j\}_{j=1}^m$ from sample trajectories with different initial conditions
2: Apply generator EDMD (Algorithm 1) to obtain sKO eigenfunctions $\{\phi_i(x)\}_{i=1}^N$ and eigenvalues $\{\lambda_i\}_{i=1}^N$. Alternatively, for linear systems, OU eigenfunctions are computed exactly.
3: Approximate $f(x) \approx \tilde{f}(x) = \sum_{i=1}^N f_i \phi_i(x)$ via regression
4: If necessary, increase f_1 so that $\tilde{f}(x_j) > 0$ for all j.
5: Approximate solution to KBE is $\tilde{\Phi}(t, x) = \sum_{i=1}^N f_i e^{\lambda_i (T-t)} \phi_i(x)$
6: Approximate Doob transform (biasing) is $\tilde{u}(t, x) = c\mathbf{B}(x)^* \nabla \log \tilde{\Phi}(t, x)$. Choose c such that a sufficient number of trajectories reach the rare event.

2.4 Numerical examples

We demonstrate our framework on a series of linear and nonlinear stochastic dynamical systems. The impact of numerical parameters used to construct the biasing is first explored through a simple example involving a one-dimensional Ornstein–Uhlenbeck (OU) process. We then demonstrate the generality of our approach for linear dynamical systems with additive noise, by applying it to a non-normal linear SDE, a noisy

Brownian oscillator, and the stochastic advection-diffusion equation (which is an infinite-dimensional system). We then turn to several nonlinear SDEs, where we show how the approach enables escape from different types of attractors.

The stochastic ODE systems are integrated numerically using a stochastic Runge–Kutta scheme [105, 102]. The stochastic PDE system is integrated using exponential Euler methods [62]. Since our importance sampling is unbiased, it suffices to report the variance of the importance sampling weight as seen in (2.12). Without loss of generality, we will also report the relative error defined in (2.10) with $M = 1$, i.e., the *relative error per sample*.

2.4.1 Illustrative one-dimensional SDE

We first consider a simple one-dimensional OU process to illustrate our approach and to highlight numerical challenges that occur in more complex examples as well. Let $X_t \in \mathbb{R}$ evolve according to

$$\begin{cases} \mathrm{d}X_t = -X_t \,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \\ X_0 = 0. \end{cases} \tag{2.29}$$

Our goal is to estimate $\rho = \mathbb{P}(X_T \geq 2 | X_0 = 0) = \mathbb{E}[\mathbb{1}_{x>2}(X_T) | X_0 = 0]$, where $T = 1$. For this problem, the marginal density at time T can be derived analytically and the exact value of ρ (to five digits) is 1.5745×10^{-2}.

The infinitesimal generator of the system, which is the same as the stochastic Koopman generator, is

$$\mathcal{A}\psi = -x\psi' + \psi'' \tag{2.30}$$

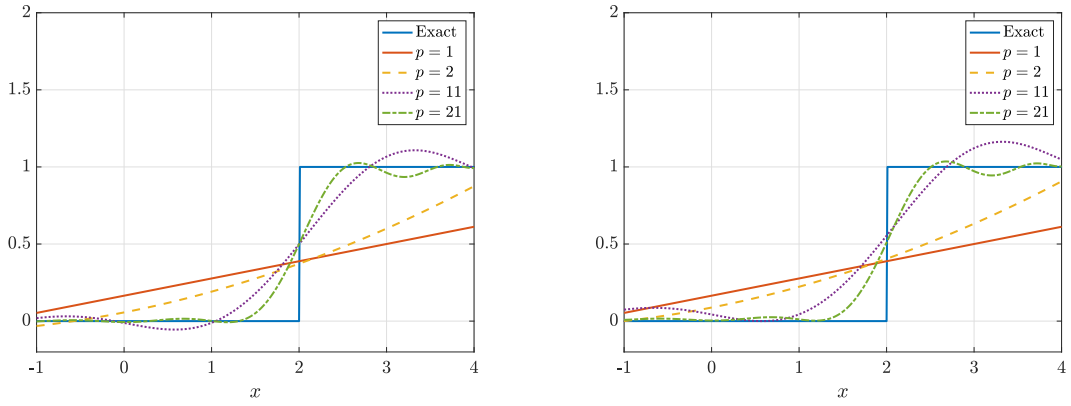
and the associated eigenvalue problem is known as the Hermite differential equation, whose solutions can be found in closed form. The eigenfunctions are the probabilists' Hermite polynomials $\phi_n(x) = \mathrm{He}_n(x)$ with eigenvalues $\lambda_n = -n$ for $n \in \mathbb{N}$. We use least squares regression to find the expansion coefficients in (2.20). In this case, the

eigenfunctions are orthogonal with respect to the standard Gaussian distribution, which implies that an optimal approximation of the indicator function $f(x) = \mathbb{1}_{x>2}(x)$ in a weighted $L^2(\nu)$ (where ν is the standard Gaussian measure) sense could be found by integrating the product of the indicator and an eigenfunction over the standard Gaussian measure. More generally, if the diffusion is reversible, then its eigenfunctions will be orthogonal with respect to the invariant distribution of the system [90]. However, this approach is impractical for higher-dimensional systems, as it would require computing several high dimensional integrals, each of which is sensitive to the rare event. Therefore, to keep this example consistent with the results of the more complicated systems, we perform regression as described in Section 2.3.4.

We perform our regression with test points drawn from a distribution with more probability mass in the rare event than the invariant distribution. Specifically, we draw $m = 50$ independent samples $x_i \sim \mathcal{N}(0, 2^2)$, where roughly 15% of points fall inside the region of interest. To mitigate the Gibbs phenomenon, we use a mollified version of the indicator in the regression problem, $f(x) = \frac{1}{2}(1 + \tanh(3(x - 2)))$. The resulting approximations, for polynomial degrees $p = N - 1 = 1$, 2, 11, 21, are plotted in Figure 2-1a. Notice that least-squares regression often leads to the approximating function not being strictly positive over the domain. As explained in Section 2.3.4, we then add a constant to the approximation such that it is strictly positive. We show the resulting approximations of the indicator function in Figure 2-1b.

Now we use the approximated observable $\tilde{f}$ to build an approximate Doob transform (2.23) and perform importance sampling via (2.13) and (2.14). To account for a diminished biasing magnitude due to positivization, as discussed in Section 2.3.4, we multiply the biasing function by a factor $c \geq 1$ to ensure that a sufficient number of trajectories reach the rare event. Below, we will explore how the choice of c impacts the performance of the importance sampling estimator.

First, Figure 2-2 shows the time-T marginal distributions of the biased and unbiased systems, along with the optimal (zero-variance) importance sampling distribution for the expectation of interest. Notice that as the number of eigenfunctions increases, the shape of the histogram tends towards the zero-variance importance sampling



(a) Without positivity constraint.

(b) With positivity constraint.

Figure 2-1: Approximating the indicator function.

density. Table 2.1 reports the variance and relative error per sample of the importance sampling estimators resulting from each approximation of f. In this simple example, we see that increasing the number of basis functions does not meaningfully increase the efficiency of the estimator. This is likely because increasing the polynomial degree of the approximation leads to more local minima and maxima, causing some sample trajectories to be driven away from the rare event of interest. On the other hand, this result demonstrates how even a small number of eigenfunctions can significantly improve the efficiency of importance sampling. For instance, using just two eigenfunctions results in the variance being reduced by a factor of 20 compared to simple Monte Carlo.

	Variance	Relative error
Monte Carlo	1.62×10^{-2}	8.07
IS $p = 1$	6.89×10^{-4}	1.67
IS $p = 2$	7.62×10^{-4}	1.75
IS $p = 11$	5.56×10^{-4}	1.50
IS $p = 21$	2.84×10^{-4}	1.07

$\rho_{\text{true}} = 1.57 \times 10^{-2}$

Table 2.1: One-dimensional OU example: IS estimator variance with increasing polynomial degree p. The multiplier c and offset ε are tuned.

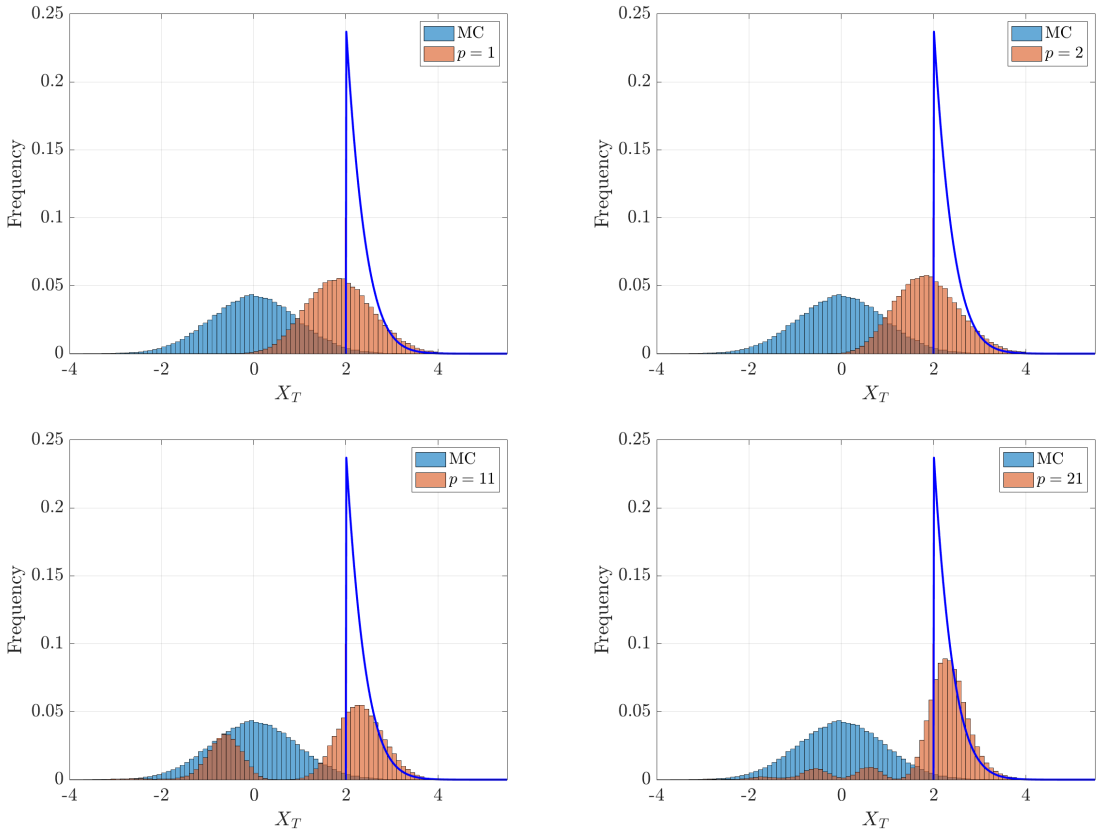


Figure 2-2: Sample distribution at time $T = 1$ for the one-dimensional OU example. Blue curve is the optimal importance sampling density.

Ideally, the multiplier c should be chosen so that the variance of resulting importance sampling estimator is as small as possible. In Table 2.2 we demonstrate the impact of c on this variance, fixing $p = 1$. The variance of the estimator initially decreases with increasing c, up to a threshold beyond which the performance degrades. Intuitively, too small a multiplier results in a larger variance, as too few samples reach the rare event. On the other hand, biasing with a very large multiplier leads to many samples deep in the tails. This implies that the most of resulting weights (2.14) will be small. Since the estimator is unbiased, however, a few samples will have very large weights, leading to a large estimator variance overall. In the following numerical examples, we choose c such that roughly 50% of the resulting samples enter the rare event; this rule of thumb is justified by the trends observed in Table 2.2. If, however, most of the weights resulting from a given value of c are very small, and certainly if c is so large such that the relative error is larger than that of simple Monte Carlo, then the value of c should be reduced so that trajectories are not pushed as deeply into the tails.

	Variance	Relative error	Proportion in rare event
Monte Carlo	1.62×10^{-2}	8.07	0.0157
IS $c = 1$	6.29×10^{-3}	5.05	0.0432
IS $c = 2$	2.69×10^{-3}	3.31	0.0912
IS $c = 4$	8.76×10^{-4}	1.89	0.284
IS $c = 6$	7.88×10^{-4}	1.79	0.558
IS $c = 8$	4.27×10^{-3}	4.16	0.789
IS $c = 16$	3.64×10^{-1}	38.4	0.999

$\rho_{\text{true}} = 1.57 \times 10^{-2}$

Table 2.2: One-dimensional OU example: impact of the multiplier on importance sampling performance, with fixed $p = 1$. Rightmost column reports the proportion of sample trajectories terminating in the rare event region.

2.4.2 Linear examples

We now consider linear SDEs of the form,

$$\mathrm{d}X_t = \mathbf{A}X_t\,\mathrm{d}t + \mathbf{B}\,\mathrm{d}W_t, \tag{2.31}$$

where $\mathbf{A} \in \mathbb{R}^{d\times d}$ is diagonalizable, has eigenvalues $\{-\sigma_i\}_{i=1}^d$ with strictly negative real parts, and right and left unit eigenvectors denoted by $\{e_i\}_{i=1}^d$ and $\{w_i\}_{i=1}^d$, respectively. Here $\mathbf{B} \in \mathbb{R}^{d\times r}$ and $W_t \in \mathbb{R}^r$ is an r-dimensional Brownian motion. We also assume that the left eigenvectors of $\mathbf{A}$ are not in the null space of $\mathbf{B}$.

For linear stochastic dynamical systems, the sKO eigenfunctions can be found exactly. The generator of the sKO semigroup for linear SDEs is known as the Ornstein–Uhlenbeck (OU) operator. In was shown in [82] that, under mild conditions, the OU operator has a discrete spectrum in $L^p(\nu)$, where ν is the stationary measure of the process. Furthermore, [82] shows that the eigenfunctions are complete in $L^p(\nu)$ for $p \geq 2$, they have a polynomial structure, and the eigenvalues and eigenfunctions are the same for all p. Computing the eigenfunctions, however, presents a separate challenge. It is well known that if the OU operator is self-adjoint, which is the case if $\mathbf{A}$ and $\mathbf{B}$ are symmetric and commute, then the eigenfunctions are the tensorized Hermite polynomials [90],

$$\phi_{\mathbf{n}}(x) = \prod_{k=1}^{d} \mathrm{He}_{n_i}\left(\frac{\sqrt{2\sigma_i}}{\|\mathbf{B}^* e_i\|}\langle x, e_i\rangle\right). \tag{2.32}$$

If $\mathbf{A}$ is normal with only complex eigenvalues, and $\mathbf{B}$ commutes with $\mathbf{A}$, then the eigenfunctions are a tensor product of Hermite-Laguerre-Itô polynomials, first noted in [24]. Otherwise, one has to consider numerical methods for computing the eigenfunctions [67, 71, 138].

Non-normal dynamics

We begin with a two-dimensional non-normal system, where

$$\mathbf{A} = \begin{bmatrix} -1 & 0 \\ 1 & -0.3 \end{bmatrix}, \quad \mathbf{B} = 0.1\mathbf{I}. \tag{2.33}$$

The inner product of the two eigenvectors of $\mathbf{A}$ is 0.8192, which reflects the degree of non-normality of the linear system. The eigenvalues are $\sigma_1 = -0.3$ and $\sigma_2 = -1$, with left eigenvectors $[0.8192, 0.5735]^T$ and $[1, 0]^T$, respectively. We consider the problem of escaping from a ball of radius L at a fixed finite time T,

$$\rho = \mathbb{P}[\|X_T\| \geq L \,|\, X_0 = 0], \tag{2.34}$$

where $T = 10$, $L = 0.75$. This is a problem of escaping from an attractor, which is well-studied in the computational chemistry literature. Methods such as transition path theory [121] and the string method [130] characterize the most likely pathways for trajectories to transition between metastable states. Related methods such as the gentlest ascent dynamics [131] find transition paths by pushing the system along the direction of the most slowly decaying *right* eigenvector of the Jacobian at the stable point. In a separate effort, [104] justifies using most slowly decaying right eigenvector to construct efficient importance sampling estimators for linear stochastic PDEs, in the presence of a suitable spectral gap. Yet these methods are typically restricted to gradient systems (noisy diffusions on a potential energy surface) or self-adjoint linear systems. Using the Koopman approach, we will demonstrate below that biasing a non-normal linear system along the *left* eigenvector that corresponds to the most slowly decaying mode leads to a significantly better importance sampling estimator.

For the system in (2.33), one can easily check that the first six eigenfunctions, ordered according to the magnitudes of the eigenvalues and with total polynomial

degree up to two, are

$$\begin{aligned}
\phi_0(x) &= 1, & \lambda_0 &= 0 \\
\phi_1(x) &= \sqrt{200\sigma_1}\langle x, w_1\rangle & \lambda_1 &= -0.3 \\
\phi_2(x) &= \sqrt{200\sigma_2}\langle x, w_2\rangle & \lambda_2 &= -1 \\
\phi_3(x) &= 200\sigma_1\langle x, w_1\rangle^2 - 1 & \lambda_3 &= -0.6 \\
\phi_4(x) &= 200\sqrt{\sigma_1\sigma_2}\langle x, w_1\rangle\langle x, w_2\rangle - 2\frac{\sqrt{\sigma_1\sigma_2}}{\sigma_1+\sigma_2}\langle w_1, w_2\rangle & \lambda_4 &= -1.3 \\
\phi_5(x) &= 200\sigma_2\langle x, w_2\rangle^2 - 1 & \lambda_5 &= -2
\end{aligned}$$

where w_1 and w_2 are left eigenvectors of $\mathbf{A}$. The function of interest is an indicator on the ball of radius 0.75 centered at the origin, which is projected onto the set of eigenfunctions. Since the indicator is an even function along the direction of any solitary left eigenvector, we can omit eigenfunctions with odd polynomial degree prior to projection. Thus, the indicator function over the ball is projected onto the span of $\{\phi_0, \phi_3, \phi_4, \phi_5\}$.

We plot the eigenfunctions in Figure 2-3 and highlight the left eigenvector directions in red. To generate the regression points that are used to approximate the indicator function as a linear combination of eigenfunctions, we simulate 121 independent trajectories of length T, beginning with uniformly spaced initial conditions on $[-0.8, 0.8]^2$. The state is extracted at time intervals of $\Delta t = 0.02$, and the resulting 60621 points are shown in Figure 2-3. In this example we found

$$\Phi(t,x) = 0.035 + 0.0342\phi_3(x)e^{\lambda_3(T-t)} - 0.0323\phi_4(x)e^{\lambda_4(T-t)} + 0.0092\phi_5(x)e^{\lambda_5(T-t)} \tag{2.35}$$

with multiplicative factor $c = 7$. Figure 2-4 then shows the vector fields produced by the resulting biasing function at $t \in \{5, 8, 9.9\}$. Notice that the biasing pushes in the direction of the slowest-decaying left eigenvector for most of the simulation period $[0, T]$, until the end of the interval, when $T - t$ becomes small and the biasing (2.23) begins to push in all directions away from the attracting point. In Figure 2-5, we

show sample trajectories of the unbiased and biased systems. Notice that the exit directions do not generally align with any eigenvector directions. Performance of the importance sampling estimator is summarized in Table 2.3, where we observe that the variance is reduced by a factor of over 6000. Figure 2-6 shows the histogram of the norm of the state at time T for simple Monte Carlo and importance sampling. Notice also that far more samples reach the rare event when importance sampling is applied.

To show that method works well with larger values of T, we also consider the case where $T = 50$ and apply the same biasing scheme, i.e., using (2.35) but with the new T value. For this case, the estimator performs similarly well and we see that the variance is reduced by a factor of over 3000. The results are summarized in Table 2.4. The quality is maintained mainly due to the nature of the problem we are solving. We only consider the probability that the state is in the rare event at a particular time T, rather than being in the event at any time before T. This means that the biasing function need not be very large until close to time T. This is reflected in the form of the Doob transform—the biasing function is initially small, but grows exponentially as t approaches T.

The effectiveness of biasing in the direction of the slowest-decaying left eigenvector can be explained intuitively by considering the phase portrait of a deterministic non-normal linear dynamical system. In Figure 2-7, we plot trajectories of a highly non-normal stable linear system with initial conditions on the unit circle. We also plot the left eigenvector with the least negative eigenvalue. First, note that there are initial conditions for which the norm of the state initially grows before decaying towards zero; this is a hallmark of highly non-normal systems. Second, notice that pushing outwards in the direction of the left eigenvector naturally exploits the system's transient growth to move the state even further from the attracting point at the origin. In non-normal systems, left and right eigenvectors corresponding to different eigenvalues are orthogonal. Therefore, the slowest-decaying left eigenvector is orthogonal to the (fast-decaying) manifold spanned by all but the slowest-decaying right eigenvector. When pushing in the direction of the left eigenvector, trajectories are driven away from the attracting point by the fast-decaying manifold of the dynamical system. The

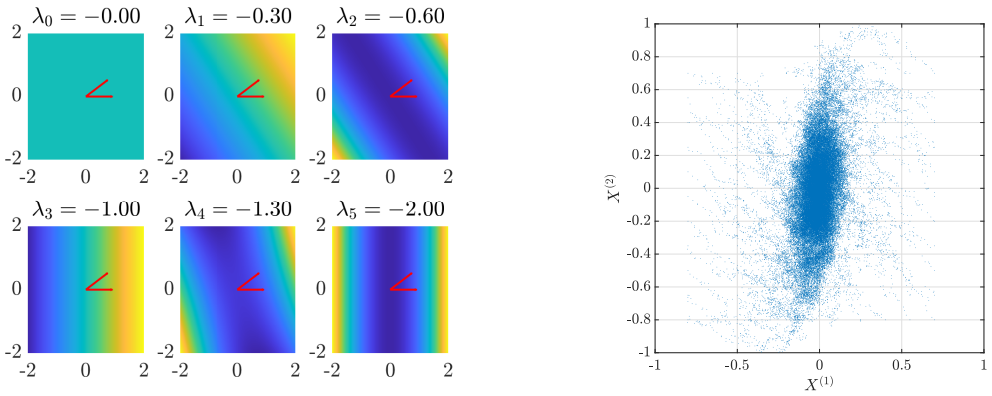


Figure 2-3: Left: Eigenfunctions of the non-normal SDE. Red vectors illustrate the left eigenvectors of $\mathbf{A}$. Right: regression points generated from sample trajectories of the unbiased system.

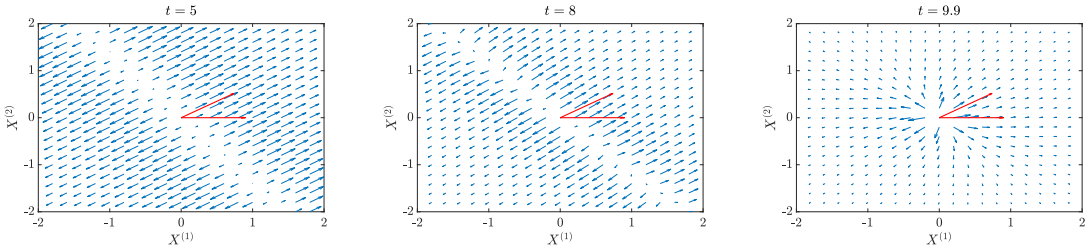


Figure 2-4: Vector fields of the biasing for the non-normal linear system at different times. Red vectors illustrate the left eigenvectors of $\mathbf{A}$. Note that the lengths of the vectors for a given time are plotted relative to each other and are not comparable for different times.

left eigenvector direction thus harnesses the system's underlying dynamics to reach the rare event region with the least biasing effort.

	Variance	Relative error
Monte Carlo	1.64×10^{-5}	246.8
Importance sampling	2.72×10^{-9}	3.18

$\rho_{\text{true}} = 1.64 \times 10^{-5}$

Table 2.3: Importance sampling performance for the SDE with non-normal dynamics. Here, $T = 10$.

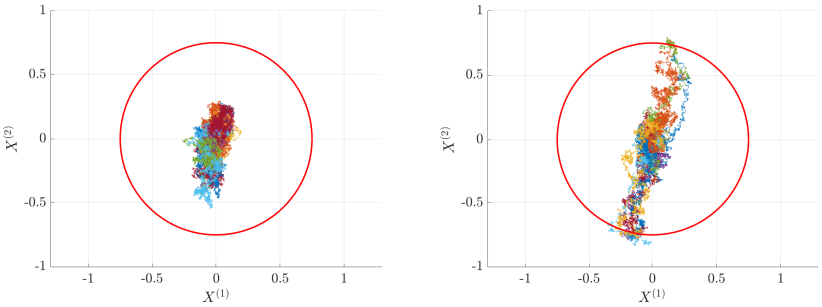


Figure 2-5: Samples of the nominal and biased trajectories of the non-normal linear system. Red circle denotes the boundary of the rare event.

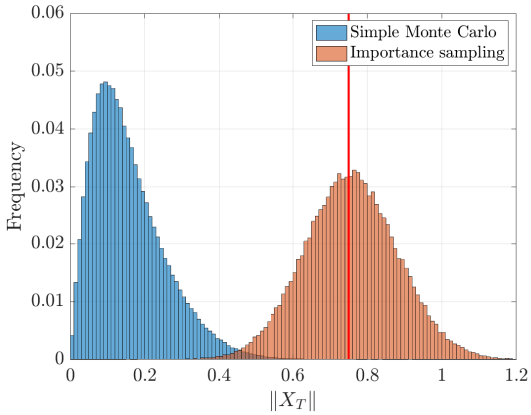


Figure 2-6: Distribution of the norm of X_T for simple Monte Carlo and importance sampling of the linear non-normal system. Red line denotes the boundary of the rare event region.

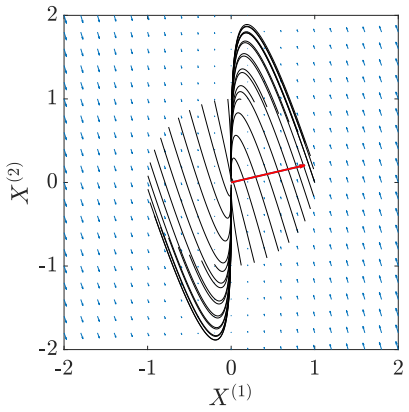


Figure 2-7: Example phase portrait of a highly non-normal system. The red vector points in the direction of the left eigenvector with the least negative eigenvalue. Notice that initial conditions that lie on the line defined by this eigenvector will initially experience transient growth before decaying to the origin.

	Variance	Relative error
Monte Carlo	1.74×10^{-5}	239.6
Importance sampling	5.60×10^{-9}	4.30

$\rho_{\text{true}} = 1.74 \times 10^{-5}$

Table 2.4: Importance sampling performance for the SDE with non-normal dynamics. Here, $T = 50$.

Stochastically-forced damped harmonic oscillator

Next we consider a damped harmonic oscillator forced by Brownian motion:

$$\begin{cases} \ddot{x} + 2\zeta\omega_0\dot{x} + \omega_0^2 x = \dot{W}_t \\ x(0) = x_0,\, \dot{x}(0) = 0\,. \end{cases} \tag{2.36}$$

This example will show that our framework works well with complex eigenvalues and *rank-deficient noise*. The oscillator can be put in the form of (2.31) with

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -\omega_0^2 & -2\zeta\omega_0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{2.37}$$

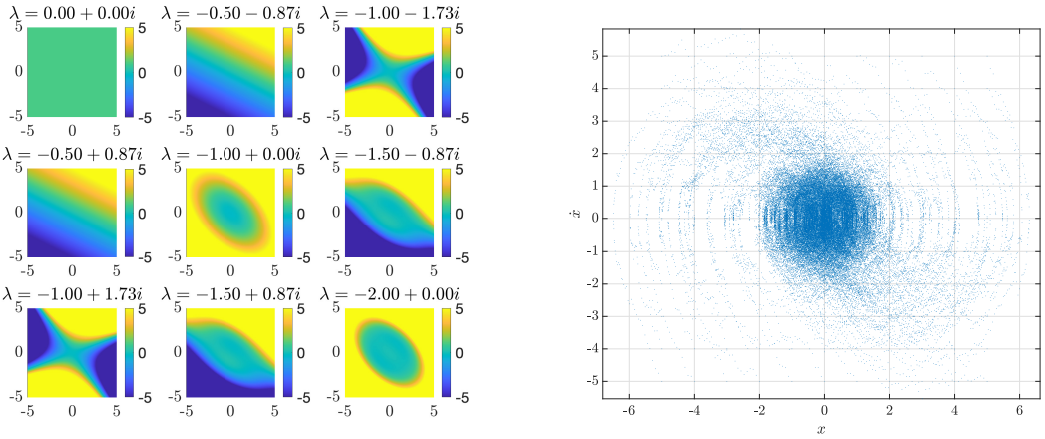


Figure 2-8: Left: Exact eigenfunctions of the Brownian oscillator. Only the real part of each eigenfunction is plotted. Right: regression points based on sample trajectories.

We compute $\mathbb{P}[|x(T)| > L \,|\, x(0) = \dot{x}(0) = 0]$, i.e., the probability that the position of the oscillator exceeds some threshold by a fixed time given that it was initially at rest. We set $\omega_0 = 1$, $\zeta = 0.5$, $L = 3$, and $T = 10$.

To the best of the authors' knowledge, for the oscillatory case, all rare event simulation algorithms require solving an associated optimal control problem. Here, we instead project an indicator function dependent on the first component of the state, $\mathbb{1}_{|x|>3}$, onto the first nine sKO eigenfunctions. The eigenfunctions are expressed as linear combinations of the Hermite–Laguerre–Itô polynomials; see [24, 138]. We plot the real parts of these eigenfunctions in Figure 2-8. Regression points are generated by simulating 121 independent trajectories of length T with uniformly spaced initial conditions on $[-5, 5]^2$ and extracting the state every $\Delta t = 0.02$ time units. In this example, the constant $c = 6$. In Figure 2-9, we show sample trajectories of the unbiased and biased systems. Notice that the impact of the biasing only seems prominent towards the end of the simulation, e.g., from $t = 8$ onward. Intuitively, this is because the system has an attracting point at zero, and since we want samples to escape at the end of the simulation, it is not advantageous to bias early.

In Figure 2-10 we show histograms of the absolute values of the position of the two systems at time T. The estimator performance is summarized in Table 2.5, where we observe that biasing reduces the variance by a factor of nearly 5000.

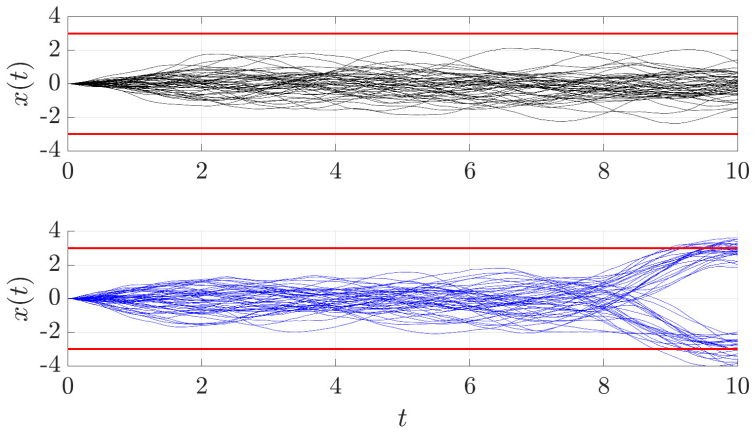


Figure 2-9: Sample paths of the unbiased and biased Brownian oscillator.

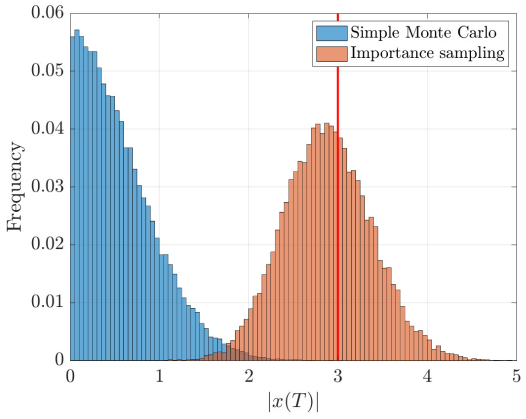


Figure 2-10: Histograms of $|x(T)|$ obtained using simple Monte Carlo and importance sampling for the Brownian oscillator.

	Variance	Relative error
Monte Carlo	2.28×10^{-5}	209.5
Importance sampling	5.10×10^{-9}	3.13

$\rho_{\text{true}} = 2.28 \times 10^{-5}$

Table 2.5: Importance sampling performance for the Brownian oscillator.

Stochastic advection-diffusion

The stochastic advection-diffusion equation is an *infinite*-dimensional non-normal linear system. We have,

$$\begin{cases} v_t = bv_x + \alpha v_{xx} + \sqrt{\epsilon}\eta \\ v(t,0) = v(t,1) = 0 \\ v(0,x) = 0, \end{cases} \tag{2.38}$$

where η is space-time white noise. Following the approach in [104], this system can be converted into the form of (2.31), where $\mathbf{A}v = bv_x + \alpha v_{xx}$ acts on the space of L^2 functions over $x \in [0,1]$ that satisfy the above boundary conditions, $\mathbf{B}$ is the identity map, and W_t is a cylindrical Wiener process. The system is discretized using an exponential Euler method [62]. Details about the numerical method used to simulate this process are described in Appendix C.2. We estimate $\mathbb{P}\big[\|v(T,\cdot)\|_{L^2([0,1])} \geq L\big]$ given that the system initially started at $v(t=0,x) = 0$. We have $b = 1$, $\alpha = 0.1$, $\epsilon = 1$, $T = 10$, and $L = 2.5$.

We compute the biasing for the system based on the SDE that arises from the discretized version of the stochastic advection-diffusion equation. In this case, we use only two eigenfunctionals: the constant functional and the second order eigenfunctional $\phi_2(v) = \sqrt{2\mu_1}\langle v, w_1\rangle^2 - 1$, where w_1 is the eigenfunction of the L^2-adjoint of $\mathbf{A}$, $-\sigma_1$ is the leading eigenvalue of $\mathbf{A}$, and $\langle u, v\rangle = \int_0^1 uv\,dx$. We measure the degree of non-normality of the system by either looking at the Péclet number, which is equal to $b/\alpha = 10$, or the inner product between the first two eigenfunctions of the advection-

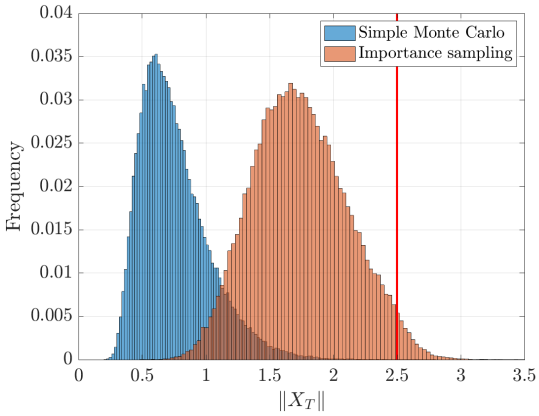


Figure 2-11: Histograms of $\|v(T,\cdot)\|_{L^2([0,1])}$ computed using simple Monte Carlo and dynamic importance sampling for the stochastic advection-diffusion equation.

diffusion operator, which is equal to 0.9147. In this example, we obtain

$$\Phi(t,v) = 0.1434\phi_2(v) + 1.1434 \tag{2.39}$$

with $c = 20$.

We plot the histogram of the norm of the system for the biased and unbiased systems in Figure 2-11, and present the results of the sampling methods in Table 2.6. Using only two eigenfunctionals, the variance of the estimator is reduced by a factor of 12 over Monte Carlo.

	Variance	Relative error
Monte Carlo	2.02×10^{-5}	222.5
Importance sampling	1.68×10^{-6}	64.24

$\rho_{\text{true}} = 2.02 \times 10^{-5}$

Table 2.6: Importance sampling performance for the stochastic advection-diffusion equation.

2.4.3 Nonlinear examples

Van der Pol oscillator

We now demonstrate our approach on nonlinear stochastic systems. Consider the noisy Van der Pol oscillator given by,

$$\mathrm{d}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ m(1-x_1^2)x_2 - x_1 \end{bmatrix} \mathrm{d}t + \sqrt{2\epsilon}\begin{bmatrix} \mathrm{d}W_1 \\ \mathrm{d}W_2 \end{bmatrix}. \tag{2.40}$$

In the absence of noise, the system exhibits a limit cycle, such that all initial conditions converge to it (except the origin, which is an unstable equilibrium). In the presence of noise, trajectories cluster on a band that is centered on the limit cycle of the deterministic system. We consider the problem of 'peeling' a solution of the stochastic system from this band. Let $m = 0.3$, $\epsilon = 0.01$, and $T = 10$; our task is to estimate

$$\mathbb{P}\big[x_1(T)^2 + x_2(T)^2 > 2.7^2 \mid x_1(0) = 2,\, x_2(0) = 0\big]. \tag{2.41}$$

The initial condition lies on the limit cycle of the deterministic system. The rare event is a region that lies outside of it.

We first find the sKO eigenfunctions of the system. As described in Section 2.3.3, we apply gEDMD, using a basis $\{\psi_k(x_1, x_2)\}_{k=1}^n$ of bivariate Legendre polynomials with total degree up to 10. This basis is constructed such that it is orthonormal with respect to the uniform measure on $\mathcal{D} = [-4, 4]^2 \subset \mathbb{R}^2$. There are $n = 66$ elements in this basis. We generate test points by using trajectory data beginning at 400 initial conditions uniformly spaced on $\mathcal{D}$. Each trajectory is simulated on the interval $t \in [0, 10]$. The test points are generated by sampling the trajectories at intervals of $\Delta t = 0.05$, for a total of 8×10^4 test points.

We find that the quality of the eigenvalues and eigenfunctions obtained via gEDMD is highly sensitive to the polynomial degree and the choice of basis. Indeed, it is well-noted that EDMD methods can often lead to spurious eigenvalues, i.e., eigenvalues that are non-physical, when the choice of basis is poor [16, 67]. We find that the same

is true for *eigenfunctions* obtained via gEDMD, when either the basis is not sufficiently representative of the eigenfunctions or the test points do not sufficiently cover the state space. Obtaining good approximation of the eigenfunctions is critical to our sampling framework. Given a set of candidate eigenfunctions produced by gEDMD, we cross-validate them with an independent dataset generated in the same fashion as the test points. In particular, the mean-square error of a candidate eigenfunction $\phi(x)$ with eigenvalue λ is defined as $\frac{1}{m}\sum_{i=1}^{m}|\mathcal{A}\phi(x_i) - \lambda\phi(x_i)|^2$. Only candidate eigenfunctions with a testing error below some threshold (here chosen to be 0.04) are used to approximate the Doob transform. In Figure 2-12 we show the first nine approximated (and validated) sKO eigenfunctions, alongside a scatterplot of the test points.

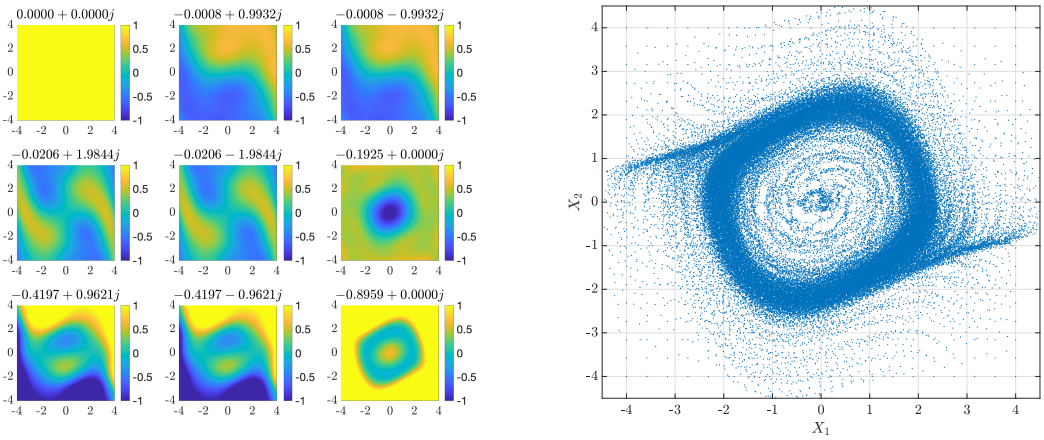


Figure 2-12: On the left, the first nine stochastic Koopman eigenfunctions for the Van der Pol oscillator. Eigenfunctions are ordered according to the magnitude of the real part of the Koopman eigenvalues, and only the real part of each eigenfunction is plotted. Right figure shows the test points.

Approximating the Doob transform to estimate (2.41) requires approximating the indicator function over the rare event region in the sKO eigenbasis. We first express the indicator function in the Legendre basis by solving a least-squares problem on the gEDMD test points. Since the Koopman eigenfunctions are approximated in the same basis, we can immediately compute the coefficients of the indicator's sKO eigenfunction expansion. Just as in the linear case, if the expansion in the sKO

eigenfunction basis is negative in some region of the domain of interest, we add a constant so that the approximation to $f(x)$ has value greater than 0.01. A scaling factor of $c = 17$ is again applied to the biasing so that enough samples will reach the rare event. In this example, we use fifteen eigenfunctions, of which nine are plotted in Figure 2-12, to approximate the Doob transform.

In Figure 2-13, we show 25 unbiased and biased sample paths of the oscillator. Notice that none of these unbiased sample paths reaches the rare event—they all remain inside the red circle demarcating the rare event region—while many of the biased paths do reach it. In Figure 2-14, we show the histogram of norm of the state at time $T = 10$ for the two systems. We report simulation results for the estimators in Table 2.7, and observe that the importance sampling estimator reduces variance by a factor of more than 400.

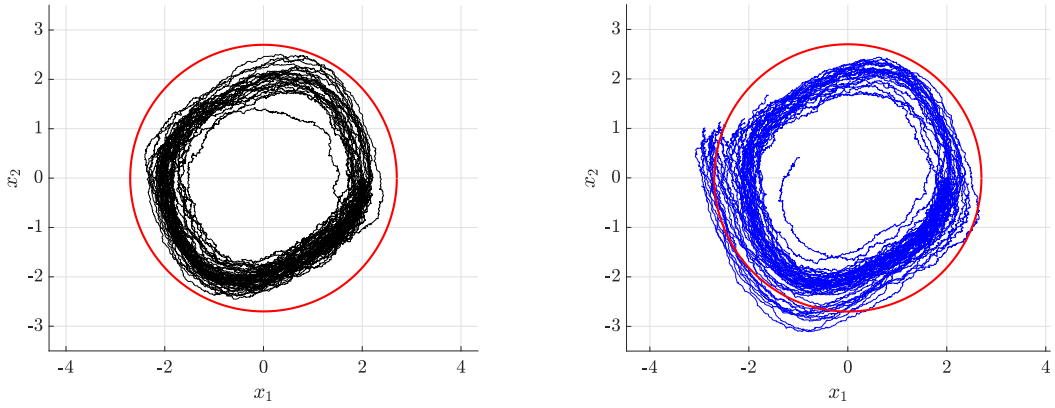


Figure 2-13: Left: sample paths of unbiased Van der Pol oscillator. Right: sample paths of biased Van der Pol oscillator. Red circle denotes boundary of the rare event.

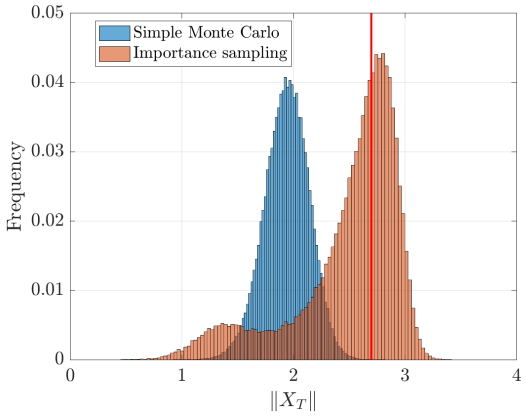


Figure 2-14: Distribution of norm of Van der Pol state at time $T = 10$. Red line denotes boundary of the rare event.

	Variance	Relative error
Monte Carlo	1.69×10^{-5}	243.01
Importance sampling	4.03×10^{-8}	11.85

$\rho_{\text{true}} = 1.69 \times 10^{-5}$

Table 2.7: Importance sampling performance for the Van der Pol oscillator.

Duffing oscillator

Now we consider the noisy Duffing oscillator,

$$\ddot{x} + \delta\dot{x} + x(\beta + \alpha x^2) = \sqrt{2\epsilon}\,\dot{W}_t,$$

which can be rewritten in standardized form as

$$\mathrm{d}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ -\delta x_2 - x_1(\beta + \alpha x_1^2) \end{bmatrix} \mathrm{d}t + \sqrt{2\epsilon}\begin{bmatrix} 0 \\ \mathrm{d}W_t \end{bmatrix}. \tag{2.42}$$

The deterministic Duffing oscillator has three equilibria. The origin is an unstable equilibrium, while $x^* = \pm\sqrt{-\beta/\alpha}$ are two stable equilibria. In the basins of attraction of the stable equilibria, the system exhibits damped oscillatory dynamics. In the stochastic setting, noise can infrequently cause trajectories to transition between the

basins of attraction. For a transition to occur, the stochastic forcing must "kick" the system in the correct direction and with the correct magnitude, in critical regions of the state space. We thus consider the rare event of *transitioning* from one basin of attraction to the other:

$$\mathbb{P}[x_1(T) > 0 \,|\, x_1(0) = -1.5,\, x_2(0) = 0].$$

Here we use parameter values $\alpha = 1$, $\beta = -1$, $\delta = 0.5$, $\epsilon = 0.0025$, and $T = 10$. The study of noise-induced transitions between attractors in dynamical systems is an important problem that arises in protein folding and chemical kinetics [121, 122, 27].

Similar to the Van der Pol oscillator, we find the stochastic Koopman eigenfunctions by applying gEDMD with a basis of bivariate scaled Legendre polynomials of total degree up to 12. This leads to 91 basis functions. To create test points for gEDMD, we simulate 400 independent trajectories over the interval $t \in [0, 10]$, with initial conditions uniformly spaced over $\mathcal{D} = [-2.5, 2.5]^2$. The data set is then generated by sampling each trajectory at intervals of $\Delta t = 0.2$. In Figure 2-15, we show the first nine approximated and validated eigenfunctions of the stochastic Duffing oscillator, along with the scatter plot of the test points. We approximate the indicator function $f(x) = \mathbb{1}_{x_1>0}(x_1, x_2)$ via a linear combination of these nine sKO eigenfunctions, using regression on the same test points. As before, we add a constant to the approximation so that the minimum value of the approximation to $f(x)$ is greater than 0.01, and scale the biasing term with a multiplicative factor $c = 8$.

In Figure 2-16, we show 25 of the resulting biased sample trajectories of the Duffing oscillator, compared to unbiased paths. The few unbiased trajectories shown here do not transition to the opposite basin of attraction. We plot a histogram of the final positions of the unbiased and biased sample trajectories in Figure 2-17. The figure demonstrates that unlike simple Monte Carlo, the biased trajectories are able to sample the transition paths with much greater success. Quantitative performance of the estimators is compared in Table 2.8. In particular, it can be seen that the importance sampling estimator reduces variance by a factor of nearly 5000.

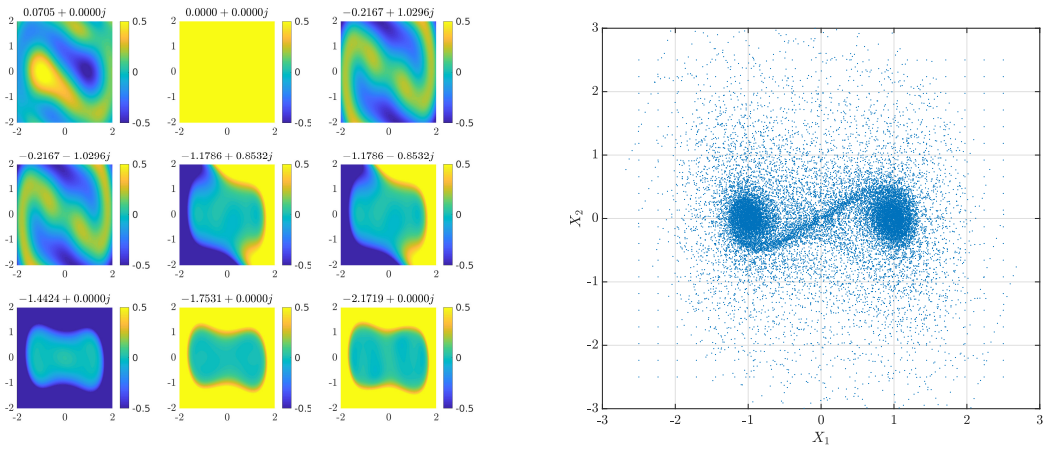


Figure 2-15: First nine stochastic Koopman eigenfunctions of the noisy Duffing oscillator. Eigenfunctions are ordered according to the magnitude of the real parts of the Koopman eigenvalues.

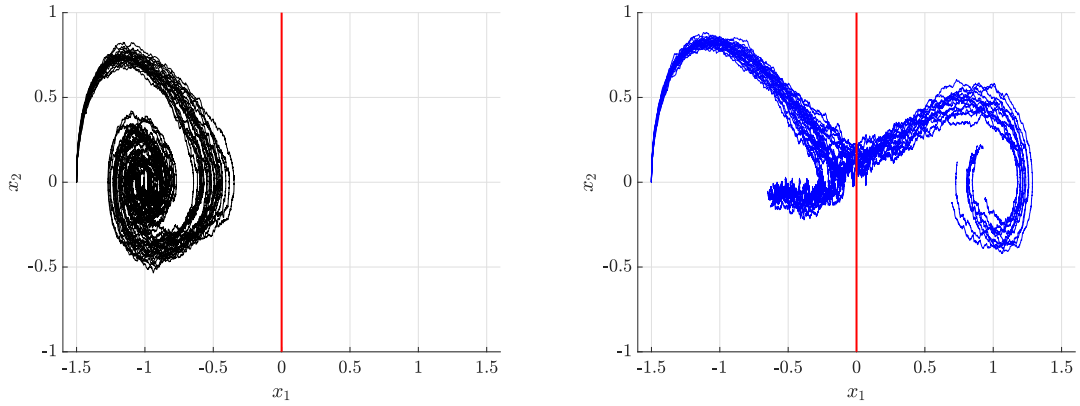


Figure 2-16: Left: sample paths of the unbiased Duffing oscillator. Right: sample paths of the biased Duffing oscillator. Red line denotes the boundary of the rare event.

	Variance	Relative error
Monte Carlo	2.11×10^{-5}	217.93
Importance sampling	4.35×10^{-9}	3.13

$\rho_{\text{true}} = 2.11 \times 10^{-5}$

Table 2.8: Importance sampling performance for the noisy Duffing oscillator.

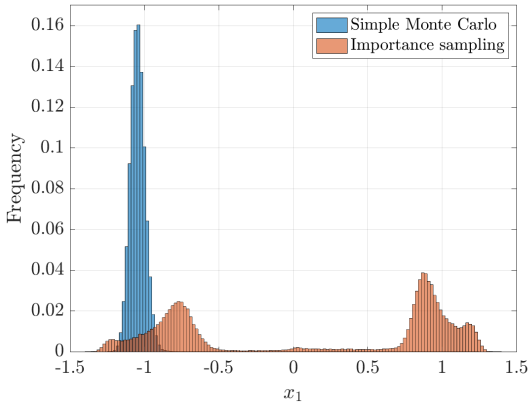


Figure 2-17: Noisy Duffing oscillator: histogram of x_1 at time $T = 10$ for the unbiased and biased systems.

2.5 Analyzing the second moment

It is useful to understand how the approximation of the Doob transform impacts the variance of the resulting importance sampling estimator. Here we provide some simple analytical results to that end. For this analysis, we assume that the sKO eigenfunctions are obtained exactly. Consequently, the only error in the solution to the KBE originates from the accuracy of approximation of the terminal condition.

We perform a non-asymptotic analysis of the importance sampling scheme based on the approach outlined in [41, 40]. Assume $f(x) \geq 0$ and define $h(x) = -\log f(x)$; recall that $f(x)$ is the terminal condition of the KBE in (2.5), which is typically the indicator function over the rare event. In contrast to our earlier presentation of the Doob transform, here f is allowed to be a true indicator function, rather than a mollified version of it. Indeed, the analysis in [41, 40] takes this scenario into account. Recall that the importance sampling estimator (cf. (2.11)) of $\rho = \mathbb{E}^{0,x_0}[f(X_T)]$ can be written as

$$\Gamma(x_0) = e^{-h(\tilde{X}_T)} \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(\tilde{X}),$$

where x_0 is the initial condition. The second moment of the importance sampling

estimator corresponding to any control $u(t,x)$ in the SDE system (2.13) is

$$Q(x_0;u) = \mathbb{E}_{\mathbb{Q}}\left[e^{-2h(\tilde{X}_T)}\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)^2\right].$$

Using [13] and the subsequent analysis in [40, 41], we obtain the following variational representation of the second moment of the importance sampling estimator:

$$-\log Q(x_0;u) = \inf_{v\in\mathcal{V}} \mathbb{E}\left[\frac{1}{2}\int_0^T \|v(s)\|^2\,\mathrm{d}s - \int_0^T \|u(s,\hat{X}_s)\|^2\,\mathrm{d}s + 2h(\hat{X}_T)\right], \tag{2.43}$$

where $\hat{X}$ solves

$$\begin{cases} \mathrm{d}\hat{X}_s = \left[\mathbf{A}(\hat{X}_s) - \mathbf{B}(\hat{X}_s)u(s,\hat{X}_s) + \mathbf{B}(\hat{X}_s)v(s)\right]\mathrm{d}s + \mathbf{B}(\hat{X}_s)\mathrm{d}W_s \\ \hat{X}_0 = x_0 \end{cases},$$

and $\mathcal{V}$ is the set of progressively measurable admissible processes. Recall that when f is the indicator function over set E, $h(\hat{X}_T)$ is infinity if $\hat{X}_T$ does not enter E and zero otherwise. We can, therefore, restrict the set of admissible processes so that $\mathcal{V}$ only contains controls ensuring $\hat{X}_T \in E$ with probability one. Then by Lemma A.1 in [41], we have that for any sufficiently regular functions $Z(t,x)$ and $U(t,x)$, where the control is $u = -B^*\nabla U$,

$$\begin{aligned} -\log Q(x_0;u) \geq \inf_{v\in\mathcal{V}} 2U(0,x_0) - 2\,\mathbb{E}[U(T,\hat{X})] + 2\int_0^T \mathcal{G}[Z](s,\hat{X})\,\mathrm{d}s \\ - \int_0^T \|B^*(\nabla Z - \nabla U)\|^2\,\mathrm{d}s\,, \end{aligned} \tag{2.44}$$

where $\mathcal{G}[Z] = \partial_t Z + \langle A(x), \nabla Z\rangle - \frac{1}{2}\|B^*\nabla Z\|^2 + \frac{1}{2}\mathrm{Tr}BB^*\nabla^2 Z$. The operator $\mathcal{G}$ can be obtained from the partial differential operator of the KBE, $\partial_t[\cdot] + \mathcal{A}[\cdot]$, via a change of variables $Z = -\log\Phi$.

In our approach, the controller is derived from an approximation to the solution of the KBE: $\tilde{u}(t,x) = \nabla\log\tilde{\Phi}(t,x)$. Therefore, if we choose $U(t,x) = -\log\tilde{\Phi}(t,x)$, then we can use (2.44). Recall that we have assumed $\tilde{\Phi}$ (2.22) to be constructed with the exact sKO eigenfunctions. Therefore, it is an exact solution of the KBE of the system

for $t \in [0, T)$, but does not match the terminal condition at $t = T$. Nonetheless, we have $\mathcal{G}[U] = 0$ exactly. Taking $Z = U$ then gives

$$-\log Q(x_0; u) \geq 2U(0, x_0) - 2 \sup_{v \in \mathcal{V}} \mathbb{E}\Big[U(T, \hat{X}_T)\Big]. \tag{2.45}$$

Note that the above bound is tight if $\tilde{\Phi}(t, x)$ in fact exactly matches the terminal condition, $\tilde{\Phi}(T, x) = f(x)$. In this case, we have that $U(T, \hat{X}_T) = h(\hat{X}_T)$, which, in turn, implies that the right-hand side equals $2U(t, x) = -2 \log \rho$ and therefore $Q(x_0; u) \leq \rho^2$. Since $Q(x_0; u) \geq \rho^2$ by Jensen's inequality, we conclude $Q(x_0; u) = \rho^2$. In other words, the variance of the estimator is zero.

On the other hand, when the biasing is imperfect but based on the true sKO eigenfunctions, (2.45) implies that the bound on the second moment depends on the accuracy of the approximations of $f(x)$ and of the KBE solution at the initial condition x_0 (i.e., the quantity of interest $\rho = \Phi(0, x_0)$) using the eigenfunctions. Recalling that $U(T, \hat{X}_T) = -\log \tilde{\Phi}(T, \hat{X}_T)$, observe that

$$-2 \sup_{v \in \mathcal{V}} \mathbb{E}\Big[U(T, \hat{X}_T)\Big] = 2 \inf_{v \in \mathcal{V}} \mathbb{E}\Big[\log \tilde{\Phi}(T, \hat{X}_T)\Big].$$

Appealing to the properties of the expectation and the fact that $\hat{X}_T \in E$ with probability one,

$$\mathbb{E}\Big[\log \tilde{\Phi}(T, \hat{X}_T)\Big] \geq \inf_{y \in E} \mathbb{E}\Big[\log \tilde{\Phi}(T, y)\Big] = \inf_{y \in E} \log \tilde{\Phi}(T, y).$$

Then (2.45) can be bounded from below as follows:

$$-\log Q(x_0; u) \geq -2 \log \tilde{\Phi}(0, x_0) + 2 \inf_{y \in E} \log \tilde{\Phi}(T, y). \tag{2.46}$$

This relation is an upper bound for the second moment for the importance sampling estimator. The first term reflects how well the solution approximates the quantity of interest ρ. The second term reflects how well the approximate KBE solution approximates the terminal condition in the rare event. It is important to emphasize

that these two terms are coupled to one another since the solution at the initial condition is dependent on how well the terminal condition is approximated.

Further refinement of these bounds is difficult. The framework we have presented is rather general, in the sense that we did not make strong assumptions on the properties of the stochastic dynamical system. Moreover, without prescribing closed-form or otherwise very specific approaches to positivization or scaling (i.e., choosing ε and c), it is difficult to characterize precisely how well the sKO eigenfunctions approximate the solution to the KBE. For specific classes of dynamical systems, one might be able to elucidate these bounds further, but since the emphasis of this chapter has been on a generally applicable computational approach, we leave such analyses to future work.

2.6 Discussion

In this chapter, we presented a framework for constructing importance sampling schemes for stochastic dynamical systems, using eigenfunctions of the associated stochastic Koopman operator (sKO). We use sKO eigenfunctions to approximate the Doob transform for the observable of interest, which in turn yields an approximation of the corresponding zero-variance importance sampling estimator. Our approach is broadly applicable, and we demonstrate the computation of rare event probabilities in a wide variety of linear and nonlinear SDEs. These numerical examples highlight how one can exploit non-rare (bulk) trajectories of the dynamical system to inform biasing strategies for rare event simulation. For systems where the sKO eigenfunctions cannot be derived analytically, we used generator EDMD to approximate them numerically. Our approach is agnostic to the numerical method used to approximate the sKO eigenfunctions, however, and thus as state-of-the-art methods for numerical approximation of the Koopman operator improve, our framework too will improve in accuracy and efficacy. Moreover, even imprecise applications of our approach can still lead to significant variance reduction. We demonstrate that crude approximations to the Doob transform, using only a few numerically-approximated eigenfunctions, can lead to variance reduction of several orders of magnitude over simple Monte Carlo.

We note that our approach is applicable to a wide range of stochastic dynamical systems, including many that are not typically handled by existing rare event simulation methods. Methods inspired by computational chemistry, for example, typically consider high-dimensional diffusion processes governed by a potential, i.e., gradient systems. We instead propose a single framework that enables rare event simulation in systems with non-normal dynamics, oscillatory behavior, limit cycles, and degenerate noise—which appear in a variety of scientific and engineering settings [30, 111, 135]. This framework often "recovers" solutions proposed for specific cases. For instance, in non-normal linear systems, we find that the dominant direction of biasing away from an attracting point is aligned with the leading left eigenvectors of the drift term. This is consistent with the rigorous theoretical results of [104] for infinite-dimensional *self-adjoint* linear systems, where the left and right eigendirections coincide; there, the authors found that (given a sufficient spectral gap) the best way to escape from an attractor is again to bias in the direction of the most slowly decaying eigenmode. Another interesting feature of our approach is that seems to work well even for large time horizons. Further investigation will be needed, but this property is promising as large deviations-based sampling methods often degrade with larger time horizons [41].

We can also contrast our approach with rare event simulation methods based on stochastic optimal control [57, 56, 139]. The goal of these efforts is the same as ours: to find a controller for the dynamical system that approximates the zero-variance importance sampling estimator. However, the stochastic optimal control formulation requires solving optimization problems or the associated nonlinear Hamilton-Jacobi-Bellman equation, both of which may be intractable in high dimensions. These methods attempt to *precisely* compute the Doob transform *locally*, depending on where trajectories lie in state space. In contrast, we consider the Kolmogorov backward equation, which, due to linearity, enables efficient computation based on eigenfunction information. Our approach thus *crudely* computes the Doob transform *globally*, using sKO eigenfunctions approximated via non-rare trajectories.

There are several avenues for future work. For example, approximation of the terminal condition of the KBE via sKO eigenfunctions presents some outstanding

questions. We currently construct this approximation by combining regression with a post hoc numerical correction to ensure positivity. A single integrated, consistent procedure for constructing positive approximations would be preferable: not only might it improve the efficiency of rare event simulation, but it could also enable further theoretical analysis of approximation error and hence estimator variance.

A practical bottleneck of our framework is the accuracy to which DMD methods can approximate the sKO eigenfunctions in regimes where the amount of data is limited relative to the dimensionality of the problem. Addressing this issue will be useful for scaling our approach to more complex high-dimensional systems. The resolution depends in part on how Koopman numerical methods develop in the future. Our current approach requires the ability to evaluate gradients of eigenfunctions anywhere in the state space. State-of-the-art Koopman numerical methods for high-dimensional systems such as Hankel DMD only give values of the eigenfunctions at the test points. The key to addressing these problems will be to find an alternative to importance sampling that uses only the eigenfunctions, and not their gradients, for sampling. Moreover, importance sampling is known to be, in many cases, an unstable method, as there might be no guarantees that the variance of the resulting estimator is finite [17]. The main bottleneck to how this approach scales in high dimensions for nonlinear problems is how well the stochastic Koopman eigenfunctions can be learned from data. We discuss further limitations of state-of-the-art Koopman numerical methods in Chapter 6.

In the next chapter, we present our approach to a more robust sampling methods for rare event simulation: multilevel splitting. The connection between efficient importance sampling estimators and multilevel splitting has been well established [17]. Multilevel splitting also has the virtue of being non-intrusive, meaning that one is not required to alter the system dynamics to perform rare event simulation. Since Koopman numerical methods enable us to construct crude approximations to the KBE non-intrusively, combining these methods with multilevel splitting will lead to more efficient *black box* approaches for rare event simulation.

Chapter 3

Multilevel splitting with stochastic Koopman eigenfunctions

3.1 Introduction

In the previous chapter, we showed how zero-variance importance sampling estimators for SDEs can be constructed using eigenfunctions of the stochastic Koopman operator which are computed via dynamic mode decomposition algorithms. While there exists a zero-variance importance sampling estimator, approximating it via eigenfunctions has a few drawbacks. First, importance sampling generally lacks robustness: it is possible for an IS estimator to have an arbitrarily large second moment, which implies that it can perform worse than standard Monte Carlo when the estimator is constructed poorly. Second, in the context of our Koopman framework, importance sampling requires the ability to evaluate gradients of the eigenfunctions everywhere in the state space, and many state-of-the-art DMD methods for high dimensional systems only produce eigenfunction values at the training points. Lastly, from the previous chapter, we had to appeal to *ad hoc* approaches to correct approximations to the indicator function when they evaluated to be less than zero in order to apply the Doob transform properly. For these reasons, we appeal to multilevel splitting algorithms, which addresses some these issues. The goal of this chapter is to show that the Koopman framework in the last chapter can be used to create efficient multilevel

splitting estimators.

3.2 The mechanics of multilevel splitting

We describe the basic mechanics of multilevel splitting as well as some state-of-the-art approaches for constructing efficient estimators. The problem setting of this chapter is identical to that of Chapter 2.1.1. Multilevel splitting was first presented in the nuclear engineering literature by Kahn & Harris [64]. The basic idea of multilevel splitting is to partition the state space into a sequence of nested sets with the smallest set being exactly equal to the rare event of interest. The probability of interest is then a product of conditional probabilities, each of which is not small. Each conditional probability is estimated sequentially by simulating a branching process generated from the stochastic differential equation. When a trajectory crosses into the next nested set, it is allowed to split into independent trajectories. This increases the chance that a trajectory will reach the rare event of interest. Each of these trajectories are weighted down according to how many times it has been split. In Figure 3-1, we provide a simplified diagram showing how a single initial trajectory produces a branching process that leads to a trajectory reaching the set of interest.

The rare probability of interest can then be expressed as a product of conditional probabilities, each of which is not small. Suppose we have $E = D = D_M \subset D_{M-1} \subset \cdots \subset D_1 \subset D_0$. Then we have that

$$\mathbb{P}(E) = \mathbb{P}(D_0) \prod_{i=1}^{M-1} \mathbb{P}(D_{i+1}|D_i). \tag{3.1}$$

Basic implementations of multilevel splitting require three design parameters: 1) the number levels (or nested sets) J, the splitting rates at each level $\{R_j\}_{j=1}^{J}$, and the locations of the nested sets. Given these three parameters, the estimator for (3.1) is found as follows. Let N_j be the number of particles reaching level j with N_0 being the initial number of particles. Let R_{j-1} be the splitting rate at level $j-1$, that is,

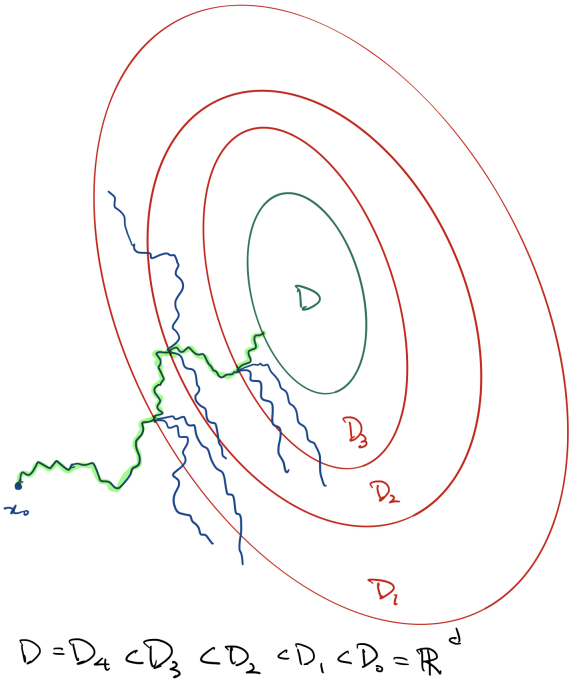


Figure 3-1: Multilevel splitting.

every particle that reaches level $j-1$ splits into R_{j-1} trajectories. Then we have

$$\mathbb{P}(D_j|D_{j-1}) \approx \frac{N_j}{R_{j-1}N_{j-1}}$$

The estimator can then be computed as

$$\hat{\rho} = \prod_{j=1}^{J} \mathbb{P}(D_j|D_{j-1}) = \prod_{j=1}^{J} \frac{N_j}{R_{j-1}N_{j-1}} = \frac{N_J}{N_0 \prod_{j=1}^{J} R_{j-1}}. \tag{3.2}$$

The locations of nested sets are described by their boundaries, which are defined as level sets of the *importance function* $U(x)$ [17, 35]. The importance function is sometimes also referred to as the *score function* [14, 20]. The partitions in state space are defined as level sets of the importance function. For example, $D_j = \{U(x) \geq d_j\}$. With this parametrization, which set the state of a trajectory belongs to can be determined solely by the evaluating the importance function. In summary, generic implementations of multilevel splitting require the definition of the following parameters:

- Importance function $U(x)$

- Number of levels J
- Locations of levels $\{d_j\}_{j=1}^J$, with $D_J = E$
- Splitting rates$\{R_j\}_{j=1}^J$.

In Algorithm 3, we outline the the basic multilevel splitting procedure as presented in [23].

Algorithm 3: Basic multilevel splitting

Input: SDE $\mathrm{d}X_t = \mathbf{A}(X_t)\mathrm{d}t + \mathbf{B}(X_t)\mathrm{d}W_t$, Importance function $U(x)$, N_0 initial particles, splitting rate $\{R_j\}_{j=1}^J$, splitting levels $\{d_j\}_{j=1}^J$.

Output: Rare event probability estimate $\hat{\rho}$

for $j = 1$ to J **do**
  for $i = 1$ to N_{j-1} **do**
    Run trajectory i until next level $U(X_t^i) > d_j$ or until final simulation time T.
  end for
  Discard trajectories that did not reach d_j
  Clone each trajectory that reach level d_j into R_j independent trajectories.
  N_j is the total number of active trajectories
end for
for $i = 1$ to N_J **do**
  Run trajectory i until final time T
end for
Compute $\hat{\rho} = \frac{1}{N_0}\frac{1}{\prod_{j=1}^J R_j}\sum_{i=1}^{N_J} \mathbb{1}_E(X_T^i)$

Naturally, the main research question here is how does one choose these algorithm parameters. The number of splitting levels and splitting rates are determined differently depending on the multilevel splitting strategy. The performance of the method will be greatly impacted by these design choices. We discuss some virtues of multilevel splitting over other rare event simulation methods such as importance sampling. Unlike importance sampling, the weights on each trajectory are always bounded below one, which implies that the splitting estimator always has bounded variance. In importance sampling, it is possible that to keep the estimator unbiased, there are certain weights greater than one which leads to higher variance than if simple Monte Carlo were used instead. Furthermore, multilevel splitting can be used with black box models that only allow access to trajectories of the dynamical system, rather than requiring access to

the model at each point of the simulation. This is desirable as many DMD approaches can work in purely data-driven settings.

On the other hand, when the design parameters are not chosen wisely, splitting can result in an intractable number of trajectories being simulated. One may end up constructing an estimator with variance that is similar to simple Monte Carlo while expending more computational effort. Intuitively, more levels and larger splitting rates produce more trajectories, which implies lower variance of the resulting estimator. However, too many particles can be intractable to produce if trajectories are expensive to simulate. Too few levels or small splitting rates may not result in any trajectories reaching the rare event. A poor partitioning of the state space can lead to higher variance while resulting in more computational effort. We discuss two established approaches for multilevel splitting in the next section.

3.2.1 Fixed rate splitting and large deviations theory

Like importance sampling, multilevel splitting has also been enhanced by large deviations theory. In [35], it was established that large deviations approaches to importance sampling can be adapted to multilevel splitting estimators. The basic observation is that stochastic differential equations often admit large deviations principles [37, 49, 123]. Consider the family of diffusion processes parametrized by ϵ,

$$\mathrm{d}X^\epsilon(t) = \mathbf{A}(X^\epsilon)\mathrm{d}t + \sqrt{\epsilon}\mathbf{B}(X^\epsilon)\mathrm{d}W_t. \tag{3.3}$$

Define the probability of interest $\rho^\epsilon(t,x) = \mathbb{P}(X^\epsilon \in E | X_t = x)$, which is a function of time and space. We say this system satisfies a large deviations principle if

$$\lim_{\epsilon\to 0} \epsilon \log \rho^\epsilon(t,x) = -U(t,x), \tag{3.4}$$

where $U(t,x) > 0$ is called the large deviations rate function. In contrast to importance sampling, in multilevel splitting we are mainly concerned with the number of particles being generated in the course of running the algorithm rather than only on the variance

of the estimator. Assuming we have a fixed splitting rate, i.e., $R_j = R$ for all j, a sensible goal is to choose the levels such that out of R particles starting at level $j-1$, on average, only one particle will reach level j. From [35], it was shown that when the importance function is chosen to be the large deviations rate function then the resulting class of estimators parametrized by ϵ is *asymptotically efficient*.

Definition 1. *We say an estimator $\hat{\rho}^\epsilon$ is* asymptotically *efficient if*

$$\lim_{\epsilon \to 0} \frac{\log \mathbb{E}[(\hat{\rho}^\epsilon)^2]}{2 \log \rho^\epsilon} = 0. \tag{3.5}$$

Asymptotic efficiency implies that the number of particles needed to keep the relative error constant will not grow exponentially as $\epsilon \to 0$. This, however, is a relatively weak guarantee, and does not provide any guarantees on the variance of the estimator. The number of levels, the splitting rate, and the locations of the trajectories to achieve an asymptotically efficient estimator has also been noted [17, 35]. Assuming one has access to the importance function, and assuming that we keep the splitting rate constant $R_j = R$, the levels should be chosen so that d_J is defined such that $E \subset D_J$, and the levels are spaced such that $d_j - d_{j-1} = \epsilon \log R$ [17].

In general, the large deviations rate function is difficult to obtain. The large deviations rate function can also be related to the solution of a Hamilton-Jacobi-Bellman (HJB) PDE [17, 35]. The HJB is difficult to solve, so it has been shown that *subsolutions* of the PDE will also produce estimators that are asymptotically efficient [43]. That is, if one can find a function $\widetilde{U}(t,x)$ such that

$$\begin{aligned} \frac{\partial \widetilde{U}}{\partial t} + H(x, \nabla \widetilde{U}) &\geq 0 \\ \widetilde{U}(T,x) &\leq \mathbb{1}_E(x) \cdot \infty, \end{aligned} \tag{3.6}$$

with $H(x,\alpha) = \langle \mathbf{A}(x), \alpha \rangle - \frac{1}{2} \|\mathbf{B}(x)^\top \alpha\|^2$ and $\widetilde{U}(0,x_0)$ chosen to match the value of $U(0,x_0)$, i.e., the true value of $U(0,x_0)$ at the initial conditions, then the resulting multilevel splitting estimator with $\widetilde{U}(t,x)$ will be asymptotically optimal. There is no established way of finding these subsolutions numerically; rather there are insights

in how to construct subsolutions by hand for different types of stochastic dynamical systems in [17].

The HJB equation can be related to the Kolmogorov backward equation described in Chapter 2. The KBE for the system in (3.3) is

$$\frac{\partial \Phi^\epsilon(t,x)}{\partial t} + \langle \mathbf{A}(x), \nabla \Phi^\epsilon(t,x) \rangle + \frac{\epsilon}{2} \mathrm{Tr} \mathbf{B}(x)\mathbf{B}(x)^\top \nabla^2 \Phi^\epsilon(t,x) = 0 \qquad (3.7)$$
$$\Phi^\epsilon(T,x) = \mathbb{1}_E(x).$$

Recall that the terminal condition is determined by the indicator over the rare event. By performing a variable transformation $U^\epsilon(t,x) = -\epsilon \log \Phi^\epsilon(t,x)$, we obtain a PDE for $U^\epsilon(t,x)$

$$\frac{\partial U^\epsilon(t,x)}{\partial t} + \langle \mathbf{A}(x), \nabla U^\epsilon(t,x) \rangle - \frac{1}{2} \|\nabla U^\epsilon(t,x)\|^2 + \frac{\epsilon}{2} \nabla^2 U^\epsilon(t,x) = 0 \qquad (3.8)$$
$$U^\epsilon(T,x) = \infty \cdot \mathbb{1}_E(x),$$

where we take the convention $\infty \cdot 0 = 0$. Taking ϵ to zero, we obtain (3.6). This suggests that the solution to the KBE could be used to define the importance function. We can use the same approximate solution to the KBE via stochastic Koopman eigenfunctions and use it to define the level sets for multilevel splitting instead. We demonstrate this in the numerical examples. However, we first provide some additional intuition for why the KBE solution is useful for splitting and provide another approach to choosing multilevel splitting parameters.

Intuition for why the KBE solution is optimal for fixed effort splitting

Here we give some intuition into why the solution to the Kolmogorov backward equation is optimal for defining where the splitting levels should be placed. This discussion is not a formal proof. Recall that $\rho^\epsilon(t,x) = \mathbb{P}(X_T^\epsilon \in E | X_t^\epsilon = x)$. The rare event probabilities we consider satisfy large deviations principles. We have that $\lim_{\epsilon \to 0} \epsilon \log \rho^\epsilon(t,x) = -U(t,x)$, and so informally, one may write $\rho^\epsilon(t,x) \approx \exp(-U(t,x)/\epsilon)$. Next, notice

that for $t' < t < T$,

$$\mathbb{P}(X_T^\epsilon \in E | X_{t'}^\epsilon \in D_{j-1}) = \mathbb{P}(X_T^\epsilon \in E | X_t^\epsilon \in D_j) \mathbb{P}(X_t^\epsilon \in D_j | X_{t'}^\epsilon \in D_{j-1}).$$

Then, again, informally, this implies

$$\mathbb{P}(X_t \in D_j | X_{t'} \in D_{j-1}) \approx \exp(-U(t', x_{j-1})/\epsilon + U(t, x_j)/\epsilon)$$

for some $x_{j-1} \in D_{j-1}$ and $x_j \in D_j$. Suppose we wish to have a fixed splitting rate at each level to be R and wish to have the levels so that $\mathbb{P}(D_j | D_{j-1}) \approx 1/R$, one should choose levels such that $\Delta U(t, x_j) = U(t', x_{j-1}) - U(t, x_j) = \epsilon \log R$. And therefore, we should choose the levels sets such that $D_j = \{U(t, x) \leq j\epsilon \log R\}$. Lastly, recall that $\rho^\epsilon(t, x)$ is the solution to the Kolmogorov backward equation, and therefore, we should choose $U(t, x) = -\log \rho^\epsilon(t, x)$.

3.2.2 Adaptive multilevel splitting algorithm (AMS)

Adaptive multilevel splitting is another well-established approach for choosing parameters in the MS algorithm [22]. In contrast to *fixed rate* splitting in the previous section, AMS is a *fixed effort* approach in which the number of trajectories being simulated at any given iteration is held fixed. "Poor performing" trajectories, i.e., trajectories with low importance function evaluations, are eliminated from the set of active trajectories. Rather than allowing trajectories to split each time they first enter a new subset, the splitting rate is chosen so that the number of new trajectories equals the number being culled. In effect, the splitting rate is random and fractional depending how many trajectories are eliminated during one iteration of the method. The splitting levels are found based on where the trajectories in the course of a single iteration of the algorithm. Specifically, the maximal importance function value is computed for each trajectory, and then the splitting level is based on a certain quantile of those importance function evaluations. We provide the AMS method in Algorithm 4 based on [14] and [23].

Algorithm 4: Adaptive multilevel splitting

Input: SDE $\mathrm{d}X_t = \mathbf{A}(X_t)\mathrm{d}t + \mathbf{B}(X_t)\mathrm{d}W_t$, Importance function $U(t,x)$, N_0 initial particles, maximum level $\xi_{\max}$, Resampling size K.

Output: Rare event probability $\hat{\rho}$

1: Simulate N_0 independent trajectories
2: Compute $M_i = \max_{t\in[0,T]} U(t, X_t^i)$
3: Order M_i in ascending order.
4: Current level is L, the Kth largest M_i.
5: **while** $L < \xi_{\max}$ and M_i not all equal **do**
6: Discard K_m trajectories that $M_i < L$
7: From trajectories that are not discarded, randomly draw K_m trajectories
8: Find when trajectories first reach level L and simulate K_m new trajectories at those locations. The total number of trajectories should equal N_0.
9: $\hat{\rho} \leftarrow \hat{\rho} \cdot \left(1 - \frac{K_m}{N_0}\right)$
10: **end while**
11: $\hat{\rho} \leftarrow \hat{\rho} \frac{1}{N_0} \sum_{j=1}^{N_0} \mathbb{1}(X_T^j \in E)$

While not obvious, this algorithm is unbiased [14, 23]. In contrast to fixed rate splitting, the number of iterations AMS will take is random. However it has been proven that the method will stop in a finite number of steps. In fact, it will stop, on average, in $\mathcal{O}(-N_0 \log \rho + \sqrt{N_0})$ number of steps for any importance function [21]. This implies that the number of iterations AMS takes depends linearly on the number of initial trajectories and the probability of the event. The upper limit of the importance function defines when the algorithm needs to terminate. If the algorithm only terminates on a set that is contained in the rare event of interest, then the resulting estimator will not be unbiased as it is no longer computing the correct rare event probability. In other words, the smallest set in the sequence is no longer equal to the rare event of interest. When defining the upper limit of the importance function, we need that the set E is contained in the set that is defined by the upper limit.

3.2.3 The optimal importance function

In contrast to large deviations-based fixed rate approaches, the asymptotic variance of AMS can be derived. Here we provide some theory for the optimal importance function as presented in [21, 23]. As $N_0 \to \infty$, it has been shown that AMS satisfies a central limit theorem, which has been used to show how the asymptotic variance

depends on the importance function.

Let $U(t,x)$ be the importance function and define $S_l = \inf\{s : U(s, X_s) > l\}$. Let $p_l = \mathbb{P}(S_l \leq T)$. Define the conditional probability distribution η_l to be such that when applied to a test function φ, $\eta(\varphi) = \mathbb{E}[\varphi(X_{S_l})|S_l \leq T]$, and $q^\star(x) = \mathbb{P}(X_T \in E|X_0 = x)$.

Theorem 2 ([23])**.** *The AMS estimator $\hat{\rho}$ satisfies a central limit theorem*

$$\sqrt{N}(\hat{\rho} - \rho) \longrightarrow \mathcal{N}(0, \sigma^2), \text{ in distribution,} \tag{3.9}$$

where

$$\sigma^2 = -\rho^2 \log \rho - 2 \int_{-\infty}^{\xi_{max}} Var_{\eta_l}(q^\star) p_l \, dp_l. \tag{3.10}$$

Notice that the first term is independent of the choice of the importance function. The variance is optimal when the second term is exactly zero. In [23], the authors show that when the solution to the Kolmogorov backward equation is used to define the importance function, i.e., when $U(s, X_s) = \Phi(s, X_s)$ exactly, then $\mathrm{Var}_{\eta_l}(q^\star) = 0$ identically. In fact, for AMS, any monotone increasing function of the optimal importance function is optimal [14, 23]. Unlike fixed rate splitting, AMS does not require precise values of the importance function to define the level sets, only the relative values of the importance function is important. This provides further evidence that using the solution to the KBE will lead to efficient multilevel splitting estimators. Furthermore, this theorem can give us a sense on how well an importance function is performing by comparing the empirical variance with the theoretical lower bound.

In the next section we demonstrate the use of approximate solutions to the KBE based on sKO eigenfunctions on a few canonical dynamical systems.

3.3 Numerical examples

3.3.1 Nonnormal nodal sink

We first consider a nonnormal nodal sink similar to the one in Chapter 2, except with a larger noise level. The system is

$$\begin{bmatrix} dX_1 \\ dX_2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 1 & -0.3 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} dt + \sqrt{2} \begin{bmatrix} dW_1 \\ dW_2 \end{bmatrix}$$

We estimate the probability that the system escapes a ball of radius 9 by $T = 10$: $\mathbb{P}(\|\mathbf{X}_{10}\|_2 \geq 9)$. We used five OU eigenfunctions and constructed the importance function in Figure 3-2. In Figure 3-3, we plot sample trajectories of simple Monte Carlo simulation, fixed rate splitting, and adaptive multilevel splitting. For fixed rate splitting, we studied how the variance of the estimator changes as a function of the splitting rate. Starting from a single particle, we ran each estimator 1000 times to compute its empirical variance. Then we do this procedure 10 times. To see if the extra computational cost associated with splitting is worth the variance reduction, we weigh each variance estimate by the number of times the model is evaluated. We see that when the splitting rate is only two and the number of splitting levels is held constant at five levels, the variance is worse than that of regular Monte Carlo in red. However, with a higher splitting rate, there is over variance is over 600 times smaller, and is worthwhile even with the extra computational cost incurred. We report the results in Table 3.1.

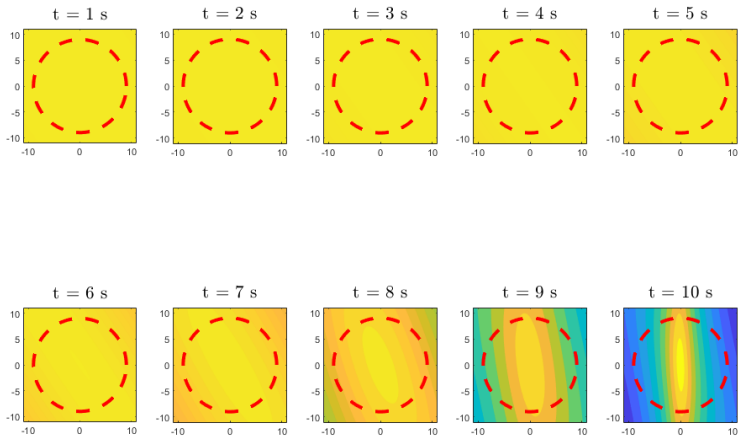


Figure 3-2: Importance function for nonnormal nodal sink.

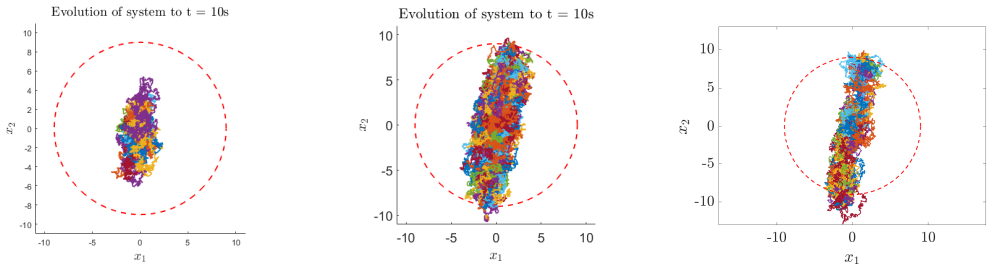


Figure 3-3: Left: Trajectories of simple Monte Carlo simulations. Middle: fixed rate splitting $N_0 = 1$. Right: Adaptive multilevel splitting $N_0 = 100$.

	Variance	Relative error per sample
Monte Carlo	2.54×10^{-4}	62.75
Fixed rate splitting (R = 2)	4.60×10^{-4}	84.46
Fixed rate splitting (R = 3)	8.19×10^{-6}	11.27
Fixed rate splitting (R = 4)	4.16×10^{-7}	2.54
AMS	1.93×10^{-6}	5.47

$\rho_{\text{true}} = 2.54 \times 10^{-4}$. Optimal AMS variance: 5.27×10^{-7}.

Table 3.1: Multilevel splitting performance for the SDE with non-normal dynamics. Here, $T = 10$.

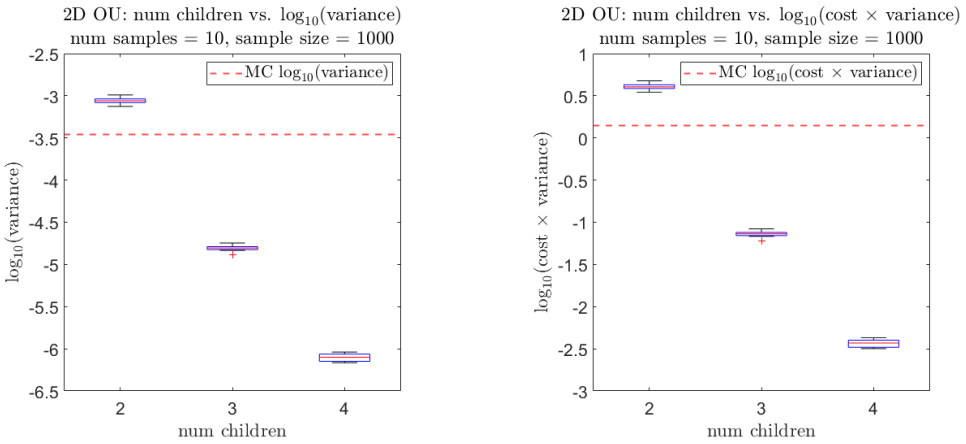


Figure 3-4: Variance and cost analysis for fixed rate splitting on nonnormal nodal sink.

3.3.2 Brownian oscillator

We study the Brownian oscillator in Chapter 2 with multilevel splitting. The system parameters and rare event are the same. We compute $\rho = \mathbb{P}(|x_{10}| \geq 3)$, and the system in state space is defined as

$$d\begin{bmatrix} x \\ \dot{x} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega_0^2 & -2\zeta\omega_0 \end{bmatrix} \begin{bmatrix} x \\ \dot{x} \end{bmatrix} dt + \begin{bmatrix} 0 \\ 1 \end{bmatrix} dW_t, \qquad \zeta = 0.5,\ \omega_0 = 1.$$

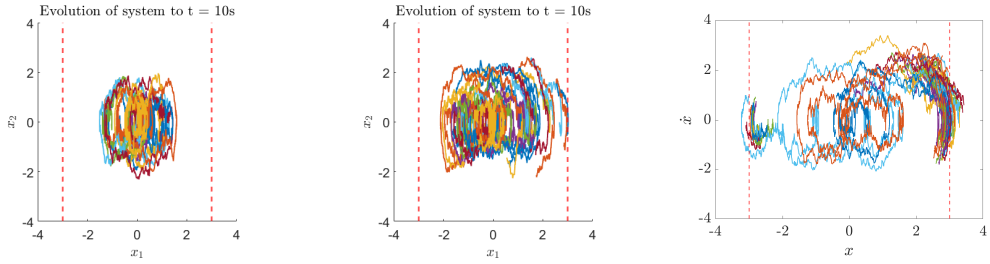


Figure 3-5: Left: Trajectories of simple Monte Carlo simulations. Middle: fixed rate splitting $N_0 = 1$. Right: Adaptive multilevel splitting $N_0 = 100$.

In Figure 3-5, we plot the trajectories of the Brownian oscillator with simple Monte Carlo simulation, fixed rate splitting, and AMS. We use the same KBE solution with nine eigenfunctions derived in Chapter 2 and constructed the importance function

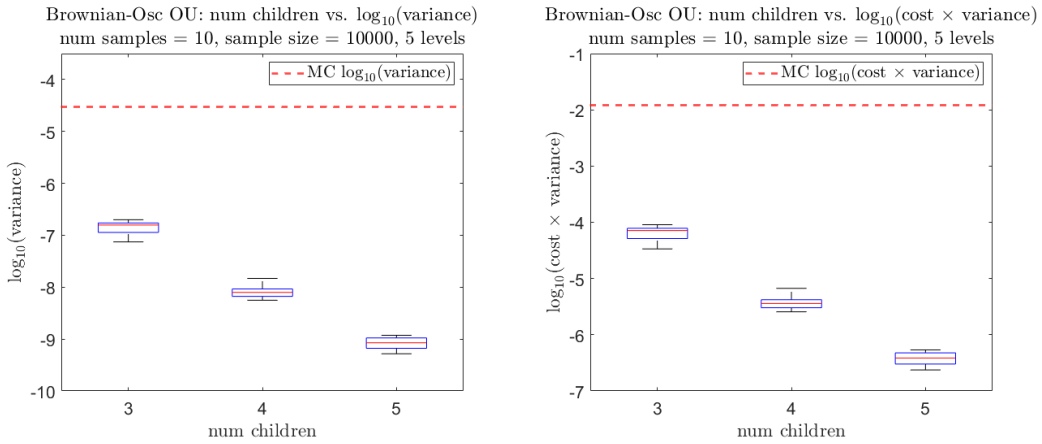


Figure 3-7: Variance and cost analysis for fixed rate splitting on the Brownian oscillator model.

in Figure 3-6. Similarly to the previous example, we study the variance reduction in fixed level splitting as a function of the splitting rate with the splitting level held fixed at five. We see substantial variance reduction over standard Monte Carlo for both fixed rate splitting and AMS, with variance being reduced by as much as 10^4 times. Full results are presented in Table 3.2.

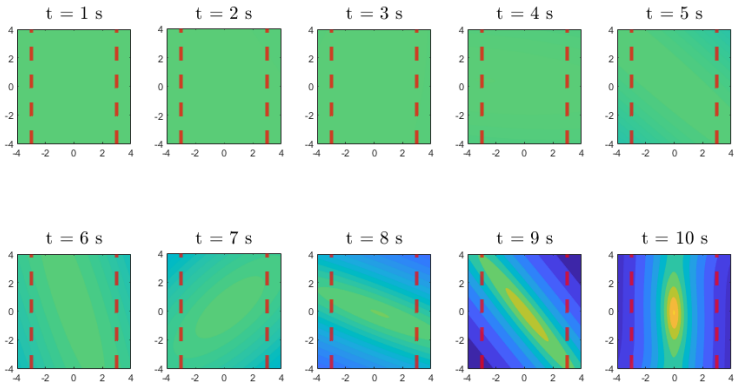


Figure 3-6: Importance function for the Brownian oscillator.

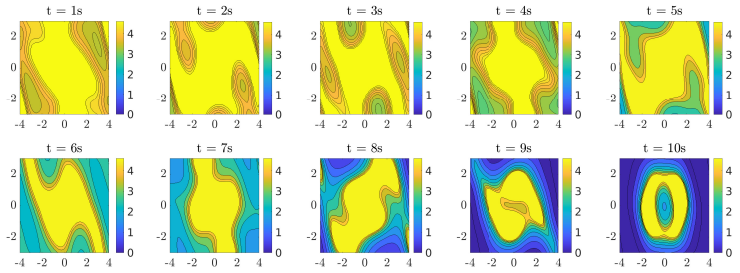


Figure 3-8: Importance function for Van der Pol oscillator

	Variance	Relative error per sample
Monte Carlo	2.88×10^{-5}	186.34
Fixed rate splitting (R = 2)	1.33×10^{-7}	12.65
Fixed rate splitting (R = 3)	7.47×10^{-9}	3.00
Fixed rate splitting (R = 4)	7.80×10^{-10}	0.97
AMS	2.28×10^{-8}	5.24

$\rho_{\text{true}} = 2.88 \times 10^{-5}$. Optimal AMS variance: 8.67×10^{-9}.

Table 3.2: Multilevel splitting performance for the Brownian oscillator. Here, $T = 10$.

3.3.3 Van der Pol oscillator

Finally, we apply AMS to the same Van der Pol oscillator as in Chapter 2. For the importance function, we use the same approximation to the KBE as in the importance sampling example. In Figure 3-8, we plot the importance function over time for the Van der Pol oscillator. In Figure 3-9, we plot fifty trajectories of the Van der Pol oscillator using AMS. Notice that not many samples are split until towards the end of the simulation. AMS reduces the variance substantially over standard Monte Carlo, and even more than importance sampling of the previous chapter.

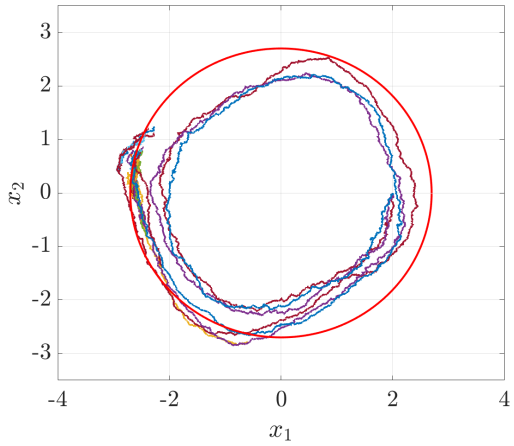


Figure 3-9: Adaptive multilevel splitting applied to the Van der Pol oscillator

	Variance	Relative error
Monte Carlo	1.69×10^{-5}	243.01
AMS	1.77×10^{-8}	7.88
Importance sampling	4.03×10^{-8}	11.85

$\rho_{\text{true}} = 1.69 \times 10^{-5}$ Optimal AMS variance: 3.14×10^{-9}.

Table 3.3: Adaptive multilevel splitting performance for the van der Pol oscillator. Importance sampling is included for comparison.

3.4 Discussion and future work

In this chapter, we considered splitting methods for rare event simulation. Using the same stochastic Koopman eigenfunctions we used for importance sampling, we found that both fixed rate splitting and adaptive multilevel splitting reduces the variance for rare event probability estimators. The approximations to the optimal importance function is just as crude as those used for importance sampling. We observed that while splitting does incur additional computational costs over standard Monte Carlo, the variance reduction makes the additional cost worthwhile. Furthermore, implementation of AMS does not require access to the drift or diffusion matrices directly, it only requires the ability to simulate trajectories of any length. This property will be

crucial when extending this framework to a fully black box formulation where the sKO eigenfunctions are found in a purely data-driven fashion. For future work, we outline a few approaches to approximating the Koopman eigenfunctions in a purely black-box manner in the next section.

3.4.1 Combining diffusion maps with stochastic Hankel DMD

We describe combining two successful approaches for accurately and robustly computing Koopman eigenfunctions with only trajectory data. The main problem with using extended dynamic mode decomposition is that it requires choosing a basis *a priori*. Choosing a basis judiciously is generally difficult to do, so methods such as the diffusion maps algorithm and stochastic Hankel DMD generate a basis through data. The diffusion maps basis is a data-driven basis constructed from the distribution of training points from the trajectory data. They are known to recover the stochastic Koopman eigenfunction for noisy gradient systems with the potential equal to the log-density of the data-generating distribution [28, 10]. Furthermore, they have also been found to be a good basis for expressing the stochastic Koopman eigenfunctions of non-gradient systems [8]. Diffusion maps is an algorithm based on kernels, which is typically limited to low dimensions. While the diffusion maps algorithm was originally conceived as means of nonlinear dimension reduction for high dimensional data, the assumption is that there exists a low dimensional manifold through which the data is fully described [29]. Stochastic dynamical systems typically do not have this property, and therefore there is generally no low dimensional manifold the test points will belong to. This makes the diffusion maps approach restricted to low dimensional systems.

Concurrently, DMD methods based on Krylov subspace methodology, such as Hankel DMD, constructs a basis through repeated action of the Koopman operator. Given an initial set basis functions, other elements of the basis are generated by applying the Koopman operator multiple times to the initial functions. The approach posits that this basis will span a Koopman invariant subspace on which eigenfunctions are defined. Here, we highlight some qualities of each approach that are attractive for multilevel splitting.

Diffusion maps approach

Diffusion maps algorithms are known for their dimension reduction abilities for high-dimensional data. They, however, have also been useful for characterizing stochastic gradient flow systems. In fact, with certain parameter choices for the diffusion maps algorithm, one can show that the eigenvectors of the stochastic matrix constructed by the algorithm is equal to the eigenfunction values of a stochastic gradient flow system with invariant measure being the data-generating distribution. This basis turns out to be, in some sense, the optimal orthonormal basis for nongradient systems. The following presentation is based mostly on [9, 10].

Suppose we are given ordered training points $\{x^{(i)}\}_{i=1}^N$ generated from a single trajectory of a stochastic dynamical system. Define the diffusion kernel

$$K_\epsilon^S(x,y) = \exp\left(-\frac{\|x-y\|^2}{4\epsilon(q_\epsilon(x)q_\epsilon(y))^\beta}\right), \tag{3.11}$$

where $q_\epsilon(x)$ is an order ϵ estimate of the sampling density $q(x) \propto \exp(-U(x))$. The following theorem shows that the diffusion kernels can be used to approximate the action of the generator corresponding to a stochastic gradient flow on functions f

Proposition 1 (From [8]). *For a function $f \in L^2(q)$ and thrice-continuously differentiable, define the functionals*

$$F_i(x_j) = \frac{K_\epsilon^S(x_i,x_j)f(x_j)}{q_\epsilon^S(x_i)^\alpha q_\epsilon^S(x_j)^\alpha}, \quad G_i(x_j) = \frac{K_\epsilon^S(x_i,x_j)}{q_\epsilon^S(x_i)^\alpha q_\epsilon^S(x_j)^\alpha}$$

where $q_\epsilon^S(x_i) = \sum_l K_\epsilon^S(x_i,x_l)/q_\epsilon(x_i)^{d\beta}$. Then the stochastic matrix

$$L_{\epsilon,\alpha,\beta}^S f(x_i) := \frac{1}{\epsilon m q_\epsilon(x_i)^{2\beta}}\left(\frac{\sum_j F_i(x_j)}{\sum_j G_i(x_j)} - f(x_i)\right) \tag{3.12}$$

$$= \hat{\mathcal{L}}f(x_i) + \mathcal{O}\left(\epsilon, \frac{q(x_i)^{(1-d\beta)/2}}{\sqrt{N}\epsilon^{2+d/4}}, \frac{\|\nabla f(x_i)\| q(x_i)^{-c_2}}{\sqrt{N}\epsilon^{1/2+d/4}}\right) \tag{3.13}$$

with high probability, where $c_1 = 2 - 2\alpha + d\beta + 2\beta$ and $c_2 = 1/2 - 2\alpha + 2d\alpha + d\beta/2 + \beta$

and $m = 2$. Here, $\hat{\mathcal{L}}$ is the generator of the system

$$dx_t = -c_1 \nabla U(x) + \sqrt{2} dW_t.$$

The eigenvectors of the stochastic matrix $L^S_{\epsilon,\alpha,\beta}$ are approximations of the eigenfunctions φ_j of $\hat{\mathcal{L}}$ evaluated at the training points. Thereafter, one way to approximate the sKO eigenfunctions is by applying EDMD with this basis directly. The reason this basis works best is because it leads to the minimal variance in the estimation of the matrix entries [9]. We propose using this basis with stochastic Hankel DMD.

Robust DMD methods

We first review the standard DMD algorithm following the presentation in [129]. Let $x_0, \ldots, x_n$ be a trajectory of the stochastic dynamical system and let $\mathbf{f}$ be a vector-valued observable and define $\mathbf{f}(i) := \mathbf{f}(x_i)$. Construct the matrices $\mathbf{X} = [\mathbf{f}(0), \ldots, \mathbf{f}(n-1)]$, $\mathbf{Y} = [\mathbf{f}(1), \ldots, \mathbf{f}(n)]$. Let $\mathbf{K} = \mathbf{Y}\mathbf{X}^\dagger$. Denote the eigenvalues of $\mathbf{K}$ to be λ_i and the left eigenvectors to be w_i. The numerical eigenvectors are $\hat{\phi}_i = w_i^T \mathbf{X}$.

If the vector of observables $\mathbf{f}$ form a Koopman invariant subspace, then, with enough trajectory data, dynamic mode decomposition produces eigenvalues that will converge to the exact Koopman eigenvalues in that subspace. Furthermore, the eigenvectors of $\mathbf{K}$ will correspond to the values of the eigenfunctions at the training points. This following result even applies when the system is stochastic.

Proposition 2 (From [129])**.** *Let $f_1, \ldots, f_k$ span a k-dimensional subspace of $L^2(\mu)$, $\mathcal{F}$, which is invariant under the action of Koopman operator $\mathcal{K}$. Let $\lambda_{j,n}$ be the dynamic eigenvalues produced by DMD along the trajectory $x_0, x_1, \ldots, x_n$. Then as $n \to \infty$, $\lambda_{j,n} \to \lambda_j$, an eigenvalue of $\mathcal{K}$ for almost every initial condition. If the eigenvalues are distinct, the numerical eigenfunctions converge to the values of the eigenfunctions along the trajectory.*

Choosing functions that span a Koopman invariant subspace, however, is generally computationally intractable. Krylov-based dynamic mode decomposition methods are better suited at approximating Koopman invariant subspaces. As the name suggests,

given a single observable, a basis of functions is construct by repeated application of the stochastic Koopman operator on that single initial observable. Just as in the finite-dimensional setting, the intuition is that a basis constructed by repeated application of the linear operator on some initial set of vectors (or basis functions) can identify invariant subspaces of the linear operator more quickly. There are two approaches that follow this intuition: the first is the *stochastic Hankel DMD*, which approximates the action of the sKO by many short term trajectories along some long single trajectory [31].

The second approach actually just uses a single long trajectory with the standard Hankel DMD to approximate the spectral quantities. This approach works for ergodic stochastic systems since the integral of the Koopman operator applied on an observable with respect to the invariant measure is equal to that of the original observable:

$$\int \mathcal{K}^t f(x) \mathrm{d}\mu(x) = \int f(x) \mathrm{d}\mu(x).$$

A guarantee of Krylov subspace-based methods working for dynamical systems is given in [129]:

Corollary 1 ([129])**.** *Let f be some observable and define the Krylov sequence $f, \mathcal{K}f, \ldots, \mathcal{K}^{k-1}f$ span a k-dimensional invariant subspace $\mathcal{F}$ and the restriction of $\mathcal{K}$ to $\mathcal{F}$ is full rank. Let*

$$\mathbf{f}(t) = [f(t)\, f(t+1)\, \ldots\, f(t+k-1)]^\top.$$

Then DMD produces eigenvalues $\lambda_{j,n}$ that converges to the eigenvalues of $\mathcal{K}$.

The practical computation of the Krylov sequences can be applied via time-delays. While Krylov DMD methods simplify the choice of basis to a set of initial observables, it still prescribe what initial observables one should choose. For future work, we propose using the diffusion maps algorithm to define a set of initial basis functions and then apply Krylov-based methods to identify a Koopman invariant subspace. We hope that the data-driven basis will provide good initial functions that when used with

robust DMD methods will produce sKO eigenfunctions that are useful for multilevel splitting.

Part II

Sampling methods *by* stochastic dynamical systems

Chapter 4

Sampling via controlled stochastic dynamical systems

4.1 Introduction

Beginning in this chapter, we shift our focus and begin exploring sampling methods *by* stochastic dynamical systems. Computing expectations with respect to high dimensional, non-Gaussian distributions is a common problem in statistics and machine learning. In this chapter, we re-interpret our approach to importance sampling for SDEs to produce controlled SDEs whose marginal at a particular time T matches target distributions.

The theory of controlled diffusions provides a different perspective of the work in Chapter 2. The Doob h-transform highlighted in that chapter exactly samples from the distribution of the time T marginal conditioned on the rare event, which is the same as the problem's zero-variance importance sampling distribution. In fact, the theory of controlled diffusions provides a unifying approach to many problems in statistics, including sampling [7], data assimilation [94], optimal transport [73], stochastic optimal control [25, 26], and rare event simulation [57, 122, 139]. Computational methods to construct the Doob h-transform efficiently have the potential to enable new approaches to all of these problems.

4.2 Controlled diffusion processes

In this section we review some relevant notions from the theories of SDEs and controlled diffusion processes [65, 85, 90, 105]. This section is similar to that of Chapter 2.2.1. Let $\{X_t\}_{t\in[0,T]}$ be a time-homogeneous d-dimensional diffusion process on $(\Omega, \mathcal{F}, \mathbb{P}, \{\mathcal{F}_t\})$. The evolution of the diffusion is described by the SDE

$$\begin{cases} \mathrm{d}X_t &= \boldsymbol{A}(X_t)\,\mathrm{d}t + \boldsymbol{B}(X_t)\,\mathrm{d}W_t \\ X_0 &= x_0 \end{cases} \tag{4.1}$$

where the drift term $\boldsymbol{A}(x)$ maps $\mathbb{R}^d$ to itself, the diffusion term $\boldsymbol{B}(x)$ maps $\mathbb{R}^d$ to the space of $d \times r$ matrices, and W_t is a standard r-dimensional Brownian motion. To guarantee existence and uniqueness of the solution, we assume that $\boldsymbol{A}$ and $\boldsymbol{B}$ are Lipschitz continuous and have linear growth. A standard tool that describes and that is used to analyze SDEs is the Markov generator defined as

$$\mathcal{A}\psi \coloneqq \sum_{i=1}^{d} \boldsymbol{A}_i(x)\frac{\partial \psi}{\partial x_i} + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}[\boldsymbol{B}(x)\boldsymbol{B}(x)^*]_{ij}\frac{\partial^2 \psi}{\partial x_i \partial x_j}. \tag{4.2}$$

This is a *linear* operator that acts on the space of twice-continuously differentiable functions and describes the evolution of expectations of the SDE through the Kolmogorov backward equation (KBE). The adjoint of the operator describes the evolution of the density of the state through the Kolmogorov forward equation, also known as the Fokker-Planck equation. Define $\eta_{t,x_0}(x)$ to be the probability density of X_t with initial condition x_0. For $f \in \mathcal{C}^2(\mathbb{R}^d)$, define $\Phi(t,x) = \mathbb{E}[f(X_T)|X_t = x]$. The Kolmogorov backward and forward equations are

$$\begin{cases} \dfrac{\partial \Phi}{\partial t} + \mathcal{A}\Phi &= 0 \\ \Phi(T,x) &= f(x) \end{cases} \qquad \begin{cases} \dfrac{\partial \eta_{t,x_0}}{\partial t} &= \mathcal{A}^*\eta_{t,x_0}(x) \\ \eta_{0,x_0}(x) &= \delta(x - x_0) \end{cases} \tag{4.3}$$

respectively, where $\mathcal{A}^*\eta \coloneqq -\nabla \cdot [\boldsymbol{A}(x)\eta] + \frac{1}{2}\operatorname{Tr}\nabla^2[\boldsymbol{B}(x)\boldsymbol{B}(x)^*\eta]$ and $\delta(x - x_0)$ is the point mass centered at x_0. Now, given an unnormalized *target density* $\pi(x)$, we wish

to find the optimal feedback control $u(t,x)$ such that the controlled diffusion process

$$\begin{cases} \mathrm{d}Y_t &= [\boldsymbol{A}(Y_t) + \boldsymbol{B}u(t,Y_t)]\,\mathrm{d}t + \boldsymbol{B}(Y_t)\,\mathrm{d}W_t \\ Y_0 &= x_0 \end{cases} \tag{4.4}$$

has its time T marginal equal to the target distribution. The control that achieves this goal is called the Doob h-transform, which we describe in the following proposition. Let $p(t,t',x,x')$ denote the Markov transition kernel of (4.1); that is, for a set A, $\mathbb{P}[X_{t'} \in A | X_t = x] = \int_A p(t,t',x,x')\,\mathrm{d}x'$. The transition kernel of the controlled process relates to that of the reference process for a certain class of controls.

Theorem 3 (Doob h-transform [38, 120, 33])**.** *Let $f \in \mathcal{C}^2(\mathbb{R}^d)$ be strictly positive over $\mathbb{R}^d$ and $\Phi(t,x) = \mathbb{E}[f(X_T)|X_t = x]$ be the solution to the KBE. Let $p(t,t',x,x')$ be the Markov transition kernel of the process in (4.1). If $u(t,x) = \boldsymbol{B}(x)^* \nabla \log \Phi(t,x)$ is the feedback control in (4.4), then the transition kernel of the controlled process is*

$$p^u(t,t',x,x') = p(t,t',x,x')\frac{\Phi(t',x')}{\Phi(t,x)}. \tag{4.5}$$

Moreover, observe that by letting $t = 0$, $t' = T$, and $x = x_0$, we have

$$p^u(0,T,x_0,x') = \eta_{T,x_0}(x')\frac{f(x')}{\Phi(0,x)} \coloneqq \eta^u_{T,x_0}(x'), \tag{4.6}$$

which is the marginal density of the controlled process at time T.

Note that this is a different version of the Doob transform as presented in Theorem 1. Now, assuming that the target distribution is absolutely continuous with respect to the marginal distribution of the reference SDE at time T, notice that if we can choose $f(x) = \pi(x)/\eta(x)$ and solve the KBE, then the corresponding Doob h-transform will lead to a controlled SDE that *samples from the target distribution.* (To compare with Chapter 2, we fixed $f(x)$ to be an indicator function over a rare event). We do not need access to the normalized target density, as the denominator in (4.6) is itself the normalization constant. This can be computed exactly if we truly had the Doob h-transform.

Suppose the initial condition $X_0 = x_0$ were not deterministic, and instead were such that $X_0, Y_0 \sim \eta_0(x)$, where $\eta_0(x)$ is any probability density on $\mathbb{R}^d$. Then the problem of finding a controller $u(t,x)$ for process $\{X_t\}$ such that $Y_0 \sim \eta_0(x)$, $Y_T \sim \pi(x)$ and such that the KL divergence from the path space measure of $\{X_t\}$ to that of $\{Y_t\}$ is minimized is known as the *Schrödinger bridge problem* (SBP) [7, 34, 91, 107, 124]. While the solution to the SBP is still in the form of a Doob h-transform, computing the solution is quite challenging and is an open area of research. State-of-the-art methods for solving the SBP rely on solving a series of optimization problems, which is quite intractable in high dimensions.

Previous work on controlled SDEs for importance sampling and rare event simulation of SDEs typically do not take this approach, citing the difficulty of solving the KBE for high-dimensional systems [122]. Instead, they consider a stochastic optimal control perspective. If we introduce the variable transform $U(t,x) = -\log \Phi(t,x)$, we obtain a stochastic Hamilton-Jacobi-Bellman equation [122], which has the following variational formulation

$$U(t,x) = \inf_u \mathbb{E}\left[\frac{1}{2}\int_t^T \|u(t,Y_t)\|^2 \mathrm{d}t - \log f(x)\middle| Y_t = x\right]. \tag{4.7}$$

This approach is common in the rare event simulation literature [57, 122, 139]. A clear proof of Proposition 3 using the stochastic optimal control formulation is provided in [120]. The stochastic control perspective has the added benefit of also describing an information-theoretic approach to the problem [57]. In the next section, we show that there exists a family of SDEs whose Doob h-transforms can be found *without* directly solving these computationally challenging problems.

4.3 Construction of the controlled SDE sampler

We construct a method for sampling a target density π on $\mathbb{R}^d$, based on finding the Doob h-transform using the KBE. Cases in which the solution to the optimal control problem can be written in closed form are rare. Yet if the reference process's Markov

generator is time-independent and has eigenfunctions, then the solution to the KBE can be written in closed form. This enables us to approximate the optimal control directly. We thus convert the optimal control problem into a static optimization problem that can be solved efficiently.

We first describe the reference process that admits closed-form expressions for the Markov generator's eigenfunctions. We then describe how we project the likelihood function described in Proposition 3 onto the basis of eigenfunctions. Next, we sketch how one can choose a good terminal marginal and what initial conditions will make computations expedient. Lastly, we describe the class of distributions to which this method applies before summarizing the approach in Algorithm 5.

4.3.1 Choosing a reference process

To make sampling via controlled SDEs tractable, we must find a way to approximate the Doob h-transform without solving a high-dimensional PDE or a series of optimal control problems. Unlike problems in dynamical systems and molecular dynamics [43, 57, 122], since we only care about sampling from some target distribution, we have freedom to choose the reference dynamical system.

In [7] the authors choose, as a reference, an overdamped Langevin system whose invariant distribution is the target distribution. Intuitively, this has the benefit that with a sufficiently large time horizon T, the control does not need to be excessive to guide the system to the invariant distribution. However, one does not know the exact density of the system at time T. Furthermore, one would still have to solve a PDE or contend with optimal control problems.

Another option is to choose the reference to be a standard Brownian motion. The density of a Brownian motion is known exactly for any time, and the solution to the KBE can be written in terms of an integral. However this integral is expensive to compute in high dimensions. This choice of reference is studied for neural SDEs in [120, 119], where it is shown that feed-forward neural networks are rich enough to approximate the Doob h-transform that will sample from a class of target distributions. These papers do not provide an algorithm for finding the control, however.

Our desiderata for a reference system are as follows: we need to be able to compute the marginal density of the uncontrolled system at time T that contains the support of the target density, and we need a way to solve the KBE or optimal control problems without expensive computations. We argue that the reference system should be a *linear* SDE, also known as an *Ornstein-Uhlenbeck* (OU) process. Specifically, the drift term should be the negative identity and the constant $d \times d$ diffusion matrix $\boldsymbol{B}$ can be user-designed depending on the target density:

$$\begin{cases} \mathrm{d}X_t &= -X_t \mathrm{d}t + \boldsymbol{B}\mathrm{d}W_t \\ X_0 &= x_0. \end{cases} \tag{4.8}$$

Like Brownian motion, if the initial condition is deterministic, the density can be derived exactly for all time. In this case, $X_t \sim \mathcal{N}(x_0 e^{-t}, \boldsymbol{\Sigma}_t)$ where $\boldsymbol{\Sigma}_t = \frac{1-e^{-2t}}{2}\boldsymbol{B}\boldsymbol{B}^*$ [65]. Furthermore, the corresponding Markov generator, called the OU operator, has a discrete spectrum and the eigenfunctions of the system can be derived exactly. The OU operator is

$$\mathcal{A}\psi(x) = -\langle x, \nabla\psi(x)\rangle + \frac{1}{2}\operatorname{Tr}\boldsymbol{B}\boldsymbol{B}^*\psi(x). \tag{4.9}$$

The eigenfunctions of this operator are products of Hermite polynomials. In particular, let $\boldsymbol{B}^* e_i = \mu_i e_i$, where $\|e_i\| = 1$, and let $\boldsymbol{n} \in \mathbb{N}_0^d$ be multi-indices. The eigenfunctions are

$$\phi_{\boldsymbol{n}}(x) = \prod_{i=1}^{d} \mathrm{He}_{n_i}\left(\frac{\sqrt{2}}{\mu_i}\langle x, e_i\rangle\right) \tag{4.10}$$

with eigenvalues $\lambda_{\boldsymbol{n}} = -\sum_{i=1}^{d} n_i$ [82]. Here, $\mathrm{He}_{n_i}(x)$ denotes the Hermite polynomial of degree n_i. This spectral decomposition lets us find an *exact* solution to the KBE as long as $f(x)$ can be expressed in terms of the eigenfunctions. Notice that if

$$f(x) = \sum_{\boldsymbol{n}} c_{\boldsymbol{n}}\phi_{\boldsymbol{n}}(x), \quad \text{then } \Phi(t,x) = \sum_{\boldsymbol{n}} c_{\boldsymbol{n}} e^{\lambda_{\boldsymbol{n}}(T-t)}\phi_{\boldsymbol{n}}(x). \tag{4.11}$$

4.3.2 Projecting onto eigenfunctions

The OU process gives us eigenfunctions that allow us to avoid costly computations associated with solving an optimal control problem or directly solving the KBE. The issue instead is that we need to find the expansion coefficients c_i for a given $f(x)$ and set of eigenfunctions $\{\phi_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathcal{I}}$. We find this "projection" by minimizing the KL divergence from the approximate density to the target.

Define $\tilde{f}(x,c) = \sum_{\boldsymbol{n}\in\mathcal{I}} c_{\boldsymbol{n}}\phi_{\boldsymbol{n}}(x)$, where $\mathcal{I} \subset \mathbb{N}_0^d$ is some set of multi-indices. Let $\pi(x)$ and $\pi_0(x)$ be the unnormalized and normalized target densities, respectively, and let $\tilde{\pi}$ be the approximate density. Let η be the density of the uncontrolled system at time T. Define $f(x) = \pi(x)/\eta(x)$ and let $\tilde{f}(x,c)$ be its approximation. Then we may write the exact and approximate densities as

$$\pi_0(x) = \frac{f(x)\eta(x)}{\gamma}, \quad \tilde{\pi}(x) = \frac{\tilde{f}(x,c)\eta(x)}{\tilde{\gamma}}, \tag{4.12}$$

where γ and $\tilde{\gamma}$ are the normalizing constants of $\pi(x)$ and $\tilde{\pi}(x)$. Here, γ is not known, but $\tilde{\gamma}$ can be computed exactly: $\tilde{\gamma} = \mathbb{E}[\tilde{f}(X_T,c)|X_0 = x_0] = \sum_{\boldsymbol{n}\in\mathcal{I}} c_{\boldsymbol{n}} e^{\lambda_i T}\phi_{\boldsymbol{n}}(x_0)$. The KL divergence from $\tilde{\pi}$ to π_0 is $\mathrm{D}_{\mathrm{KL}}(\pi_0(x)\|\tilde{\pi}(x)) = \mathbb{E}_{\pi_0}[\log\pi_0(x) - \log\tilde{\pi}(x)]$. Minimizing this divergence amounts to maximizing $\mathbb{E}_{\pi_0}[\log\tilde{\pi}]$ over the space of admissible probability densities. This objective can be directly approximated. Recall that both π_0 and $\tilde{\pi}$ can be written in terms of η, and we have

$$\max_{c\in\mathbb{R}^{|\mathcal{I}|}} \mathbb{E}_\eta\left[f(x)\log\frac{\tilde{f}(x,c)}{\tilde{\gamma}(c)}\right]. \tag{4.13}$$

Solving this optimization problem should guarantee that $\tilde{f}(x,c)$ is positive since if it were otherwise, $\tilde{\pi}$ would no longer be a density function. The objective function is approximated using Monte Carlo samples from η. Doing so enforces positivity at the sample points, since $\log\tilde{f}$ diverges otherwise; it is possible, however, that an approximation is negative elsewhere in the domain. With sufficient samples and eigenfunctions of moderate degree, we find that this issue can be avoided in practice.

4.3.3 Choosing the terminal marginal η_T and the initial condition

Next we must choose the marginal at time T, η_{T,x_0}, and design the reference OU process accordingly. Since the only closed-form solutions to the OU process are normal distributions, we restrict ourselves this class. We also assume that the initial condition is deterministic. (Choosing otherwise complicates the problem, and is closer to the full Schrödinger bridge [107].) In the optimization problem in (4.13), we are in effect evaluating the expectation $\mathbb{E}_{\pi_0}[\log \tilde{\pi}]$ using $\eta_{T,x_0} \equiv \eta$ as a biasing distribution. This means that we should choose η such that the objective function can be estimated with low variance. One way to choose η is to apply another method that would give some (crude) normal approximation to π, such as the Laplace approximation, expectation propagation, or mean-field variational Bayes. Doing so allows us to approximate the scale of the target distribution and shift it closer to the origin. Therefore, we choose the initial condition to be $x_0 = 0$.

Suppose we have in hand a normal approximation $\mathcal{N}(0, \boldsymbol{\Sigma})$ to π. To find the OU process that has this marginal at time T, we first find the eigenvalue decomposition of $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^*$. The eigenvectors identify the principal directions for the diffusion term. The diffusion matrix is then chosen to be

$$\boldsymbol{B} = \sqrt{\frac{2\boldsymbol{\Lambda}}{1 - e^{2T}}}\boldsymbol{V}. \tag{4.14}$$

4.3.4 Expressiveness of the Hermite polynomials

The OU eigenfunctions form a complete set in the weighted L^2 space with weight function equal to the invariant density η_∞ of the SDE. For example, in 1-D we need

$$\int_{\mathbb{R}} |f(x)|^2 \eta_\infty(x) \mathrm{d}x = \int_{\mathbb{R}} \frac{\pi(x)^2}{\eta_T(x)^2} \eta_\infty(x) < \infty. \tag{4.15}$$

One difficulty with having a deterministic initial condition is that the distribution of the OU process at finite time T will always be a Gaussian that is *narrower* than

the invariant distribution. This means the ratio between the invariant density and the reference marginal η_T grows as $\exp(x^2)$, which implies $\pi(x)$ must have Gaussian tails for the integral to be finite. It is possible to extend this method to include distributions with heavier tails based on the Schrödinger bridge literature [7, 91, 94], though the formulation will be more complicated.

Lastly we address the question of how to choose the set of multi-indices $\mathcal{I} \subset \mathbb{N}_0^d$. In low dimensions (approximately $d \leq 5$), it is practically feasible to use a total order basis $\|\boldsymbol{n}\|_1 \leq p$ for $p \in \mathbb{N}$. For higher dimensional problems, we will need to incorporate additional structure of the target density into the problem to reduce the number of basis functions. One option is to take advantage of the target density's Markov structure [69]. If we know that the density factors into a product of potential functions that are only dependent on a subset of the variables, we can construct a reference process whose marginal matches the structure of the target. This allows us to decouple the problem according to the Markov structure, gives us information on how to choose the basis functions, and makes the approach more scalable.

Another generally applicable choice is the sparse truncation that corresponds to choosing $\|\boldsymbol{n}\|_q \leq p$ for $q \in [0, 1)$. Doing so assumes that $f(x)$ is well approximated by polynomials that depend primarily on certain eigenvector directions, i.e., with only lower-order cross terms. This choice originates from the high-dimensional approximation literature [108].

We summarize the method in Algorithm 5.

Algorithm 5: Computing the Doob h-transform

Input: Unnormalized target density $\pi(x)$, set of multi-indices $\mathcal{I} \subset \mathbb{N}_0^d$

Output: Optimal control $u(t, x)$

1: Find an approximation $\eta(x) = \mathcal{N}(0, \boldsymbol{\Sigma})$ to $\pi(x)$, define $f(x) = \pi(x)/\eta(x)$

2: Compute $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^*$

3: Set $\boldsymbol{B} = \sqrt{\frac{2\boldsymbol{\Lambda}}{1-e^{-2T}}}\boldsymbol{V}$

4: Construct eigenfunctions $\{\phi_{\boldsymbol{n}}(x)\}_{\boldsymbol{n}\in\mathcal{I}}$

5: Draw M independent $X^{(i)} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$

6: Solve $c^* = \arg\max_{c\in\mathbb{R}^{|\mathcal{I}|}} \frac{1}{M}\sum_{i=1}^{M} f(X^{(i)}) \log \frac{\tilde{f}(X^{(i)},c)}{\tilde{\gamma}(c)}$ where
$\tilde{f}(X^{(i)}, c) = \sum_{\boldsymbol{n}\in\mathcal{I}} c_{\boldsymbol{n}}\phi_{\boldsymbol{n}}(X^{(i)})$, and $\gamma(c) = \sum_{\boldsymbol{n}\in\mathcal{I}} c_{\boldsymbol{n}} e^{\lambda_{\boldsymbol{n}}(T-t)}\phi_{\boldsymbol{n}}(x_0)$

7: Doob h-transform is $u(t, x) = \boldsymbol{B}^*\nabla \log \sum_{\boldsymbol{n}\in\mathcal{I}} c^*_{\boldsymbol{n}} e^{\lambda_{\boldsymbol{n}}(T-t)}\phi_{\boldsymbol{n}}(x)$.

After simulating the controlled SDE (4.4) independently multiple times, we can use the samples directly for approximate inference. We may also use them for importance sampling or as an independence proposal in MCMC.

4.3.5 Correcting for bias due to SDE discretization

Since we do not have the true normalizing constant of π, we can use self-normalized importance sampling [87] to compute (asymptotically) unbiased expectations, which simultaneously estimates the normalizing constant and the desired expectation. Because our samples necessarily come from the *discretization* of an SDE, however, standard methods such as the Euler-Maruyama or Milstein schemes do not truly sample from $\tilde{\pi}$; instead they sample from the approximation $\tilde{\pi}^h$ [105].[1] Because we choose to use linear SDEs as the reference process, it is possible to correct this bias.

Let η^h and $\tilde{\pi}^h$ be the densities of the time T marginals of the uncontrolled and controlled SDEs, respectively, using the Euler-Maruyama scheme with step size h. Let $N = T/h$. Then η^h is a normal distribution with mean 0 and covariance matrix

$$\Sigma_T^h = \frac{1-(1-h)^{2N}}{1-(1-h)^2} h\boldsymbol{B}\boldsymbol{B}^*. \tag{4.16}$$

[1] If the noise is additive, as in linear SDEs, the Euler-Maruyama and Milstein schemes are equivalent.

The density $\tilde{\pi}^h$ cannot be expressed in closed form, but the likelihood ratio of $\tilde{\pi}^h$ with respect to η^h can be computed. Let $\hat{Y}_k$ be the value of the discretized path at step k, and ξ_{k+1} be the sample of the standard normal random variable used in the discretization. Then

$$\log Z(Y_T) = \log \frac{\tilde{\pi}^h(Y_T)}{\eta^h(Y_T)} = \sum_{k=0}^{N-1} \langle u(t_k, \widehat{Y}_k), \xi_{k+1} \rangle \sqrt{h} + \frac{1}{2} \sum_{k=0}^{N-1} \|u(t_k, \widehat{Y}_k)\|^2 h. \tag{4.17}$$

Then observe that the following allows access to an unbiased estimator of the normalizing constant.

$$\mathbb{E}_\pi[1] = \mathbb{E}_{\tilde{\pi}^h}\left[\frac{\pi}{\tilde{\pi}^h}\right] = \mathbb{E}_{\tilde{\pi}^h}\left[\frac{\pi}{\eta^h}\frac{\eta^h}{\tilde{\pi}^h}\right] = \mathbb{E}_{\tilde{\pi}^h}\left[\frac{\pi(Y_T)}{\eta^h(Y_T)} Z(Y_T)\right]. \tag{4.18}$$

We provide details for correcting the bias as follows. We constructed controlled SDEs whose marginal at time T is $\tilde{\pi}$, an approximation to target density π. Exact samples from $\tilde{\pi}$ are obtained by taking the final positions of the simulated trajectories of the controlled SDE (4.4). However, since these samples come from the discretization of the SDE, they do not truly sample from $\tilde{\pi}$; instead they sample from $\tilde{\pi}^h$. Because we choose to use linear SDEs as the reference process, it is possible to calculate the likelihood ratio needed to correct this bias, for both importance sampling and MCMC.

Observe that the Euler-Maruyama scheme applied to (4.8) and (4.4) gives us

$$\widehat{X}_{k+1} = \widehat{X}_k - \widehat{X}_k h + \sqrt{h}\boldsymbol{B}\xi_{k+1} \tag{4.19}$$

$$\widehat{Y}_{k+1} = \widehat{Y}_k + (-\widehat{Y}_k + \boldsymbol{B}u_k)h + \sqrt{h}\boldsymbol{B}\xi_{k+1}, \tag{4.20}$$

respectively, for $k = 0, \ldots N-1$ where $h = T/N$, $t_k = kh$, $u_k = u(t_k, \widehat{Y}_k)$, and $\xi_{k+1} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. Each of the discrete-time processes above has an associated path measure, which we denote $\mathbb{Q}$ and $\mathbb{Q}^u$ respectively. Let η^h be the marginal of $\mathbb{Q}$ at time T. Note that as h approaches zero, η^h approaches the reference density η. Observe

that

$$\mathbb{E}_{\pi}[1] = \mathbb{E}_{\eta^h}\left[\frac{\pi}{\eta^h}\right] = \mathbb{E}_{\mathbb{Q}}\left[\frac{\pi(\widehat{X}_N)}{\eta^h(\widehat{X}_N)}\right] = \mathbb{E}_{\mathbb{Q}^u}\left[\frac{\pi(\widehat{Y}_N)}{\eta^h(\widehat{Y}_N)}\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{Q}^u}\right]. \tag{4.21}$$

This means we need to need to derive the density of η^h and the likelihood ratio $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{Q}^u$. The former can be derived exactly because the reference process is linear. Recall that η^h is the density of $\widehat{X}_N$. The random increments are normally distributed, so the distribution of $\widehat{X}_k$ is normal for any k. Furthermore since $\widehat{X}_0 = 0$, we simply have to find the covariance matrix of $\widehat{X}_k$. Let $\boldsymbol{\Sigma}_k = \mathbb{E}\Big[\widehat{X}_k\widehat{X}_k^*\Big]$ denote the covariance matrix of $\widehat{X}_k$. For all k we have

$$\begin{aligned}\widehat{X}_{k+1}\widehat{X}_{k+1}^* =&(1-h)^2\widehat{X}_k\widehat{X}_k^* + h\boldsymbol{B}\xi_{k+1}\xi_{k+1}^*\boldsymbol{B}^* \\ &+ (1-h)\widehat{X}_k\xi_{k+1}^*\boldsymbol{B}^*\sqrt{h} + \sqrt{h}\boldsymbol{B}\xi_{k+1}^*\widehat{X}_k^*(1-h).\end{aligned} \tag{4.22}$$

Taking the expectation and noticing that the increment ξ_{k+1} is independent of $\widehat{X}$, we have

$$\mathbb{E}\Big[\widehat{X}_{k+1}\widehat{X}_{k+1}^*\Big] =(1-h)^2\,\mathbb{E}\Big[\widehat{X}_k\widehat{X}_k^*\Big] + h\boldsymbol{B}\,\mathbb{E}\big[\xi_{k+1}\xi_{k+1}^*\big]\boldsymbol{B}^*. \tag{4.23}$$

This gives us the following recurrence relation

$$\boldsymbol{\Sigma}_{k+1} = (1-h)^2\boldsymbol{\Sigma}_k + h\boldsymbol{B}\boldsymbol{B}^* \tag{4.24}$$

which implies

$$\boldsymbol{\Sigma}_T^h := \boldsymbol{\Sigma}_N = \sum_{k=0}^{N-1}(1-h)^{2k}(h\boldsymbol{B}\boldsymbol{B}^*) = \frac{1-(1-h)^{2N}}{1-(1-h)^2}h\boldsymbol{B}\boldsymbol{B}^*. \tag{4.25}$$

To compute the likelihood ratio $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{Q}^u$, we compute the joint densities of $\mathbb{Q}$ and

$\mathbb{Q}^u$. The conditional density of $\widehat{X}_{k+1}$ given $\widehat{X}_k$ is

$$q(\widehat{X}_{k+1}|\widehat{X}_k) = \frac{1}{\sqrt{(2\pi)^d h \det \boldsymbol{B}\boldsymbol{B}^*}} \exp\left[-\frac{1}{2}(\widehat{X}_{k+1} - (1-h)\widehat{X}_k)^*(\boldsymbol{B}\boldsymbol{B}^* h)^{-1}(\widehat{X}_{k+1} - (1-h)\widehat{X}_k)\right]. \tag{4.26}$$

The joint density of $\widehat{X}_1, \ldots, \widehat{X}_N$ is then

$$q(\widehat{X}_1, \ldots, \widehat{X}_N) = \prod_{k=0}^{N-1} q(\widehat{X}_{k+1}|\widehat{X}_k). \tag{4.27}$$

In a similar fashion, the conditional density of $\widehat{Y}_{k+1}$ given $\widehat{Y}_k$ is

$$q^u\left(\widehat{Y}_{k+1}|\widehat{Y}_k\right) = \frac{1}{\sqrt{(2\pi)^d h \det \boldsymbol{B}\boldsymbol{B}^*}} \exp\left[-\frac{1}{2}(\widehat{Y}_{k+1} - (1-h)\widehat{Y}_k - \boldsymbol{B}u_k h)^*(\boldsymbol{B}\boldsymbol{B}^* h)^{-1}(\widehat{Y}_{k+1} - (1-h)\widehat{Y}_k - \boldsymbol{B}u_k h)\right] \tag{4.28}$$

and the joint density can be found similarly as in (4.27). For any k, we have

$$\begin{aligned}\log \frac{q(\widehat{Y}_{k+1}|\widehat{Y}_k)}{q^u(\widehat{Y}_{k+1}|\widehat{Y}_k)} =& -(h\boldsymbol{B}u_k)^*(\boldsymbol{B}\boldsymbol{B}^* h)^{-1}(\widehat{Y}_{k+1} - \widehat{Y}_k(1-h)) \\ &+ \frac{1}{2}(h\boldsymbol{B}u_k)^*(\boldsymbol{B}\boldsymbol{B}^* h)^{-1} h\boldsymbol{B}u_k.\end{aligned} \tag{4.29}$$

Now notice that from (4.20), $\widehat{Y}_{k+1} - \widehat{Y}_k(1-h) = h\boldsymbol{B}u_k + \sqrt{h}\boldsymbol{B}\xi_{k+1} \sim \mathcal{N}(h\boldsymbol{B}u_k, h\boldsymbol{B}\boldsymbol{B}^*)$. This allows us to write

$$\begin{aligned}\log \frac{q(\widehat{Y}_{k+1}|\widehat{Y}_k)}{q^u(\widehat{Y}_{k+1}|\widehat{Y}_k)} &= -u_k^*(hu_k + \sqrt{h}\xi_{k+1}) + \frac{h}{2}\|u_k\|^2 \\ &= -\sqrt{h}u_k^*\xi_{k+1} - \frac{h}{2}\|u_k\|^2.\end{aligned} \tag{4.30}$$

Then the likelihood ratio is

$$\log \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{Q}^u} = \log \frac{q(\widehat{Y}_1, \ldots, \widehat{Y}_N)}{q^u(\widehat{Y}_1, \ldots, \widehat{Y}_N)} = -\sqrt{h}\sum_{k=0}^{N-1}\langle u_k, \xi_{k+1}\rangle - \frac{h}{2}\sum_{k=0}^{N-1}\|u_k\|^2. \tag{4.31}$$

This formula is akin to a discrete version of the Girsanov theorem [85, 105]. With this and the expression for the density η^h, we can use (4.21) to estimate the normalizing constant. While here we have described how to construct an unbiased estimate of the normalizing constant of π, it is then simple to extend this construction to the estimation of any other expectation over π.

4.4 Numerical experiments

We test our methodology on some simple non-Gaussian distributions and Bayesian inference problems. In the 1-D and 2-D example problems, we show how the eigenfunctions approximate the target distribution, and find that only a few eigenfunctions are needed to construct a good approximation. We then apply our methodology to Bayesian logistic regression on several datasets from [52], demonstrating feasibility in higher dimensions.

In general, we have not yet prescribed a definitive way of choosing a "good" terminal reference distribution η_T. We do note, however, that the controlled SDEs seem to perform better when the ratio between the target and reference densities $f(x) = \pi(x)/\eta(x)$ grow to infinity as $|x| \to \infty$. This is due to the fact that polynomials diverge away from the origin, so they are poor at approximating functions that decay away from the origin. This implies that η should decay faster than π and is why for the 1-D and 2-D examples, the reference distribution is chosen so that the bulk of the target distribution contains the bulk of the reference.

4.4.1 One-dimensional mixture model.

We demonstrate the methodology on a 1-D Gaussian mixture model and report some trends relating to how the KL divergence converges as the eigenfunction basis is enriched. The target distribution is $\pi(x) = 0.6\mathcal{N}(x; 1.8, 0.7^2) + 0.4\mathcal{N}(x; -2.6, 0.9^2)$. We choose the reference process to be such that the marginal at time $T = 1$ is $\mathcal{N}(x; 0, 1.4^2)$. The optimization problem in (4.13) is solved with 10000 samples. In Figure 4-1, we show how the approximate density approaches the exact density as

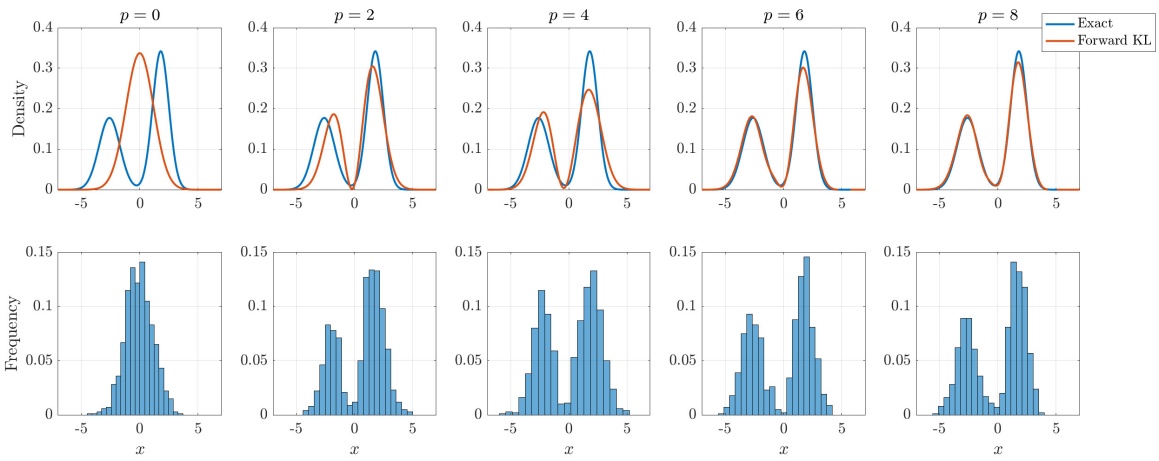


Figure 4-1: 1-D Gaussian mixture model target. Red curve is the density of the controlled SDE at $T = 1$, and p denotes the maximum polynomial order. Histograms are from samples generated by the controlled SDEs. The optimization problem in (4.13) is discretized with $M = 20000$ samples.

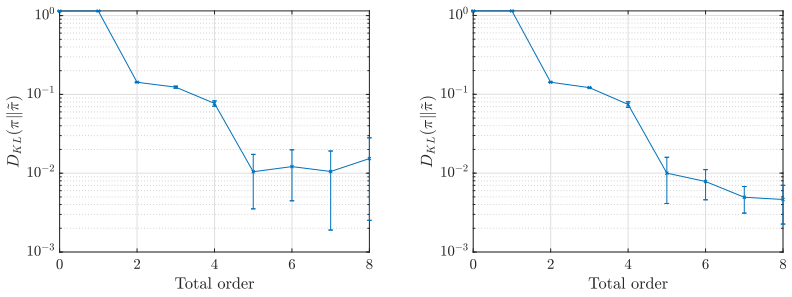
the basis is enriched as well as histograms that show samples from the approximate density. In figure 4-2, we plot how the KL divergence from the approximate to the target converges. Notice that while the KL reduces with more basis functions, the variance grows when estimating the coefficients of the higher order polynomials.

4.4.2 Two-dimensional distribution.

We evaluate the method for a highly non-Gaussian distribution. The target distribution is a slight modification of an example in [98]. The density of the two-dimensional example in Section 4.4 is $\pi(x) = \exp(-U(x))$ where

$$U(x) = 0.5\left(\frac{\|x\|_2 - 1.5}{0.7}\right)^2 - \log\left[\exp\left(-\left(\frac{x_1 - 2}{0.8\sqrt{2}}\right)^2\right) + \exp\left(-\left(\frac{x_1 + 1.5}{0.8\sqrt{2}}\right)^2\right)\right]. \tag{4.32}$$

The reference process is chosen such that the distribution at time $T = 1$ is a normal with mean 0 and covariance $\mathbf{\Sigma} = \mathrm{diag}(0.6, 1)$. We use a total order basis of degree up to p for $p \in \{0, 2, 4\}$ as shown in Figure 4-3, and see that the density is well approximated.



(a) $M = 2500$ (b) $M = 10000$

Figure 4-2: KL divergence from the approximate density $\tilde{\pi}$ to the exact target π, for repeated trials of optimization and sampling. While the divergence decreases with a richer eigenfunction basis, its variance increases. The variance decreases when more sample points are used.

h	Essential sample size ($M = 10^4$)
0.1	5.9×10^3
0.05	6.3×10^3
0.01	6.7×10^3

Table 4.1: Computed effective sample sizes for various levels of SDE discretization (time step h), using the asymptotically unbiased estimator (4.21).

In Figure 4-4, the controlled SDE is discretized with $h = 0.01$. In Table 4.1, we report the effective sample size (ESS) obtained when computing the normalizing constant $\gamma = \mathbb{E}_\pi[1]$ using the importance sampling estimator in (4.21) for various levels of discretization. For comparison, the ESS is 1.7×10^3 with $M = 10^4$ points when simply using the reference distribution η for importance sampling.

We use the samples from the $p = 4$ case to estimate normalizing constant via importance sampling. In Figure 4-4, we show that if we use $\tilde{\pi}$ directly, the resulting estimator is biased, but the mean-squared error is smaller than if we had used η as the importance sampling distribution. If we apply the bias correction, the variance of the resulting estimator increases, but it is unbiased.

4.4.3 Bayesian logistic regression.

Finally, we test our method on several Bayesian logistic regression problems, using datasets described in [52] and following the same setup. The priors on the weights $\boldsymbol{w}$

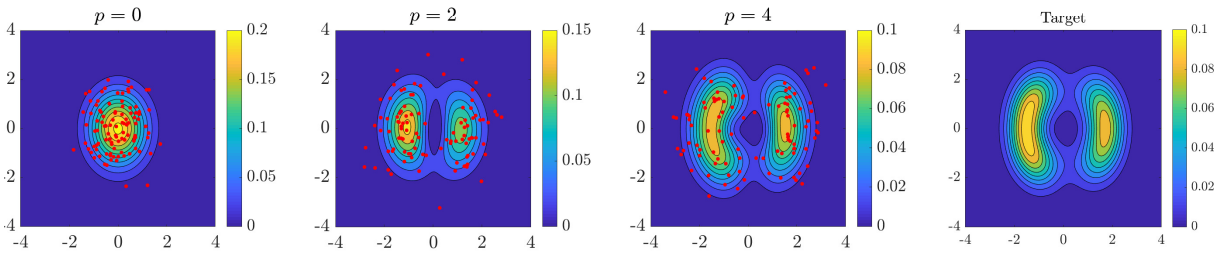


Figure 4-3: Left three figures show the approximate density produced by the controlled SDE for total degree up to p. Rightmost figure shows the exact target density. Red dots show the simulated points of the controlled SDE. The optimization problem was discretized with $M = 10000$ samples from η.

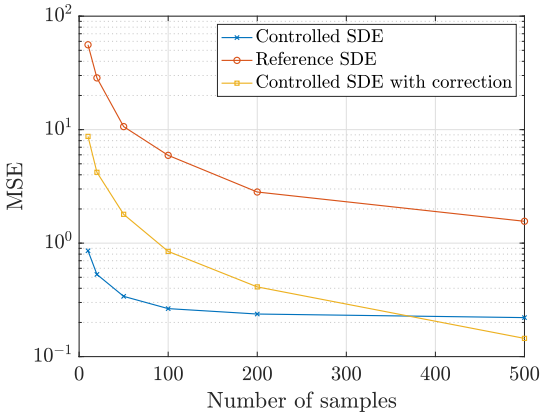


Figure 4-4: We use the $p = 4$ case as an importance sampling distribution to estimate the normalizing constant of $\pi(x)$.

Dataset	d	CSDE	NUTS
`banana`	3	0.5100	0.5094
`diabetes`	9	0.7567	0.7533
`breast_cancer`	10	0.5844	0.5844
`heart`	14	0.8100	0.8100
`thyroid`	6	0.8933	0.8800
`titanic`	4	0.7416	0.7416

Table 4.2: Testing accuracy of NUTS and controlled SDE approach from the Bayesian logistic regression datasets. (Higher numbers are better.)

are Gaussian with mean zero and covariance matrix $\alpha^{-1}\boldsymbol{I}$, and the hyperparameter α is endowed with a $\Gamma(a = 1, b = 0.01)$ distribution. We sample from the posterior distribution of $[\boldsymbol{w}, \log \alpha]$. We compare our methodology to the no U-turn sampler (NUTS) [58], and use the sample points from the controlled SDE for approximate inference. For lower dimensional problems where $d \leq 5$, we use a total order basis with order 4. For problems with $d > 5$, we use sparse truncation with $q = 1/2$. The reference process is chosen such that the marginal at time $T = 1$ is the Laplace approximation of the posterior distribution. For these examples, all SDEs are simulated with time step $h = 0.005$. In many Bayesian inference problems, evaluating the posterior is the main computational expense in sampling. To make the comparisons fair, we give both methods approximately $M = 20000$ evaluations of the log posterior or its gradient. For NUTS, this computational budget typically resulted in approximately 2000 samples, of which we use the latter 1000 to estimate the weights. For the controlled SDE case, we use the budget to solve the optimization problem in (4.13) and then generate 1000 samples of the approximate posterior using the resulting controlled SDE. In Table 4.2, we report the testing accuracy of the two methods on a variety of datasets.

4.5 Discussion and future directions

In this chapter, we introduced a tractable approach to sampling based on controlled SDEs. By choosing the reference process to be a linear SDE, we can use eigenfunctions of the system's Markov generator to approximate the required control, for any target

distribution with Gaussian tails. Future research will investigate properties of the optimization problem, expand the class of admissible distributions, and exploit low-dimensional structure or sparsity in the control. We hope that these efforts create new classes of practical methods for sampling, control, and data assimilation.

While tractable, our approach suffers from lack of robustness as small changes to the algorithm parameters can lead to schemes that fail to produce samples from the target distribution. Poor choice of the basis functions lead to lack of expressiveness so that the basis cannot properly describe the target distribution without using many more basis functions. Indeed, while we know that Hermite polynomials are able to express a broad class of likelihood ratios, in practice we found that sometimes one needs many polynomials to approximate the likelihood ratio.

In the following sections, we describe two future directions that may resolve some of the challenges preventing the framework from being more easily applicable to Bayesian computation problems.

4.5.1 Schrödinger half-bridge formulation

In the framework we described, the initial condition is always deterministic. If we appeal to the Schrödinger bridge problem (SBP) formulation, we can treat more general initial conditions, however solving SBPs generally for general initial and terminal distributions is challenging and would require solving sequences of dynamic programming problems instead [7]. Recall that our approach only requires solving a static optimization problem.

Here we consider the Schrödinger *half* bridge problem, where we allow a probabilistic initial condition, but we do not fix it to a specific distribution as in the full SBP [7, 91]. Admitting more initial conditions into the framework may lead to more to a more expressive class of approximating distributions. Following the setup before, let $\pi(x)$ be the unnormalized target density on $\mathbb{R}^d$. Let $\eta_0(x)$ and $\eta_T(x)$ be the initial

density and reference terminal density, respectively. Let the reference process be

$$\begin{cases} \mathrm{d}X_t = \mathbf{A}X_t\mathrm{d}t + \mathbf{B}\mathrm{d}W_t \\ X_0 \sim \eta_0(x). \end{cases}$$

The SBP framework requires one to find a controller $u(t,x)$ such that $Y_T \sim \pi(x)$. Since the initial distribution is no longer deterministic, the half-bridge formulation implies that the initial density η_0 will change as well. Recall that $\Phi(t,x)$ is the solution to the KBE of the reference process:

$$\begin{cases} \partial_t\Phi + \mathcal{A}\Phi &= 0 \\ \Phi(T,x) &= \dfrac{\pi(x)}{\eta_T(x)}. \end{cases}$$

Let $u(t,x) = \mathbf{B}^*\nabla \log \Phi(t,x)$, then Section 2.3 in [91] shows that the controlled process will evolve as follows

$$\begin{aligned} \mathrm{d}Y_t &= [\mathbf{A}Y_t + \mathbf{B}u(t,Y_t)]\mathrm{d}t + \mathbf{B}\mathrm{d}W_t \\ Y_0 &\sim \Phi(0,y)\eta_0(y), \end{aligned} \tag{4.33}$$

with $Y_T \sim \pi(y)$. In this chapter, we had $\eta_0(x)$ be deterministic so that the initial distribution of Y_0 was identical to that of X_0. This choice, however, is rather restrictive since η_T can only be a Gaussian with the trace of the covariance being strictly less than that of the invariant Gaussian of the reference process. We therefore would like to allow for more general initial conditions in the hope that the resulting approximating class is more expressive.

To make this framework tractable, we must be able to find initial conditions for which the Fokker-Planck is easy to solve, and the corrected initial density is simple to sample. To this end, we propose the following three classes of distributions for the initial condition:

- Discrete distributions,

- Eigenfunctions of the Fokker-Planck equation,
- Mixtures of Gaussians.

Discrete distributions

Here we consider initial densities of the form

$$\eta_0(x) = \sum_{i=1}^{N} w_i \delta(x - x_i) \tag{4.34}$$

where $\delta(x - x_i)$ is a delta distribution centered at x_i, and $w_i > 0$ are weights such that $\sum_{i=1}^{N} w_i = 1$. One can easily show that

$$\eta_T(x) = \sum_{i=1}^{N} w_i \mathcal{N}(e^{\mathbf{A}T} x_i, \Sigma_T) \tag{4.35}$$

where Σ_t solves $\dot{\Sigma}_t = 2\mathbf{A}\Sigma_t - \mathbf{B}\mathbf{B}^*$. After solving for the controller based on the KBE solution $\Phi(t, x)$, the modified initial density is

$$Y_0 \sim \Phi(0, y)\eta_0(y) = \sum_{i=1}^{N} \tilde{w}_i \delta(x - x_i). \tag{4.36}$$

The modified initial density is still discrete, but the weights are re-weighted. In this case, the modified initial condition is still simple to sample. Choosing the correct discrete distributions will require approximating the target distribution with a mixture of Gaussians with identical covariance matrices first. This approach is also a special case of the mixture of Gaussians initial condition we describe later.

Eigenfunctions of the Fokker-Planck equation

In Appendix A, we show that the eigenfunctions of the Ornstein-Uhlenbeck operator are analytically available when $\mathbf{A}$ is self-adjoint and $\mathbf{B}$ is simultaneously diagonalizable with $\mathbf{A}$. It can be easily shown that the eigenfunctions of the Fokker-Planck operator can be defined in terms of the eigenfunctions of the OU operator. Let $\eta_\infty(x)$ be the

invariant distribution, the eigenfunctions are $\phi_i(x)\eta_\infty(x)$. Suppose we have an initial density of the form

$$\eta_0(x) = \sum_{i=1}^{N} w_i \phi_i(x) \eta_\infty(x). \tag{4.37}$$

Then the time T marginal is

$$\eta_T(x) = \sum_{i=1}^{N} w_i e^{-\lambda_i T} \phi_i(x) \eta_\infty(x), \tag{4.38}$$

and the modified initial condition is

$$Y_0 \sim \left[\Phi(0, y) \sum_{i=1}^{N} w_i \phi_i(y) \right] \eta_\infty. \tag{4.39}$$

Since the KBE solution is approximated by the eigenfunctions of the OU operator, the modified initial condition is, again, a product of a polynomial and the invariant density. This density can be sampled from exactly via another controlled linear SDE whose reference samples from the invariant density exactly in finite time. We expound on this formulation in Section 4.5.2.

Mixtures of Gaussians

We can also solve the Fokker-Planck when the initial density is a mixture of Gaussians. This can be regarded as a generalization of the case with the discrete distributions. Let the initial density be

$$\eta_0(x) = \sum_{i=1}^{N} w_i \mathcal{N}(\mu_i, \Sigma_0^{(i)}). \tag{4.40}$$

The reference terminal is

$$\eta_T(x) = \sum_{i=1}^{N} w_i \mathcal{N}(\mu_i e^{\mathbf{A}T}, \Sigma_t^{(i)}) \tag{4.41}$$

where $\Sigma_t^{(i)}$ are solved via the matrix differential equation $\dot{\Sigma}_t = 2\mathbf{A}\Sigma_t - \mathbf{B}\mathbf{B}^*$ for each i. The modified initial condition is then

$$\eta_0(y) = \Phi(0, y) \sum_{i=1}^{N} w_i \mathcal{N}(\mu_i, \Sigma_0^{(i)}). \tag{4.42}$$

We would require constructing and simulating N separate controlled linear SDEs to sample from this modified initial condition.

4.5.2 A controlled SDEs formulation based on Fokker-Planck eigenfunctions

We describe in more detail the formulation for sampling via controlled SDEs using the Fokker-Planck eigenfunctions. We first find an approximation of the target density by an expansion of the eigenfunctions of the Fokker-Planck operator corresponding to some linear SDE. The eigenfunctions only depend on the covariance of the invariant measure and the eigenvectors of $\mathbf{B}$. Without loss of generality, we can choose $\mathbf{A}$ to be the identity matrix. That is, suppose we have diffusion process $\{X_t\}_{t\in[0,T]}$ evolving according to

$$\mathrm{d}X_t = -(X_t - \mu)\mathrm{d}t + \sqrt{2}\mathbf{B}\mathrm{d}W_t \tag{4.43}$$

for $\mu \in \mathbb{R}^d$, $\mathbf{B} \in \mathbb{R}^{d\times d}$, where $\mathbf{B} = \mathbf{B}^*$, and $\mathbf{B}q_k = \lambda_k q_k$, and $W_t \in \mathbb{R}^d$ is a standard d-dimensional Brownian motion. The invariant density of the system is a Gaussian with mean μ and covariance matrix $\Sigma_\infty = \mathbf{B}\mathbf{B}^*$. The generator of this process is

$$\mathcal{A}\psi = -\sum_{i=1}^{d} x_i \frac{\partial \psi}{\partial x_i} + \sum_{i,j=1}^{d} (\mathbf{B}\mathbf{B}^*)_{ij} \frac{\partial^2 \psi}{\partial x_i \partial x_j}. \tag{4.44}$$

The eigenfunctions of the generator are the tensorized Hermite polynomials:

$$\phi_{\mathbf{n}}(x) = \prod_{i=1}^{d} \mathrm{He}_{n_i}\left(\frac{\langle x - \mu, q_i\rangle}{\lambda_i}\right) \tag{4.45}$$

where $\mathbf{n} \in \mathbb{N}_0^d$ with eigenvalue $\nu_{\mathbf{n}} = -\sum_{i=1}^d n_i \lambda_i$. One can then easily check that eigenfunctions of the Fokker-Planck operator (which is the L^2 adjoint of the generator) are $\psi_{\mathbf{n}}(x) = \phi_{\mathbf{n}}(x)\eta_\infty(x)$ with the same eigenvalues. Since the eigenfunctions of the generator $\phi_{\mathbf{n}}(x)$ form an orthogonal set in $L^2(\eta_\infty)$, we have the property that $\mathbb{E}_{\eta_\infty}[\phi_{\mathbf{n}}(x)] = \int \psi_{\mathbf{n}}(x)\mathrm{d}x = 0$ for all $\mathbf{n}$.

We aim to approximate the target density by an expansion of the FP eigenfunctions, i.e.,

$$\pi_{\mathbf{w}}(x) = \sum_{\mathbf{n}} w_{\mathbf{n}} \psi_{\mathbf{n}}(x) = \left(1 + \sum_{\mathbf{n}} \mathbf{w}_{\mathbf{n}} \phi_{\mathbf{n}}(x)\right) \eta_\infty(x). \tag{4.46}$$

The constant term is always equal to one since the other eigenfunctions always integrate to zero. A feature of this parametrization is that there is no need to find a normalization constant. It is, however, possible that the density approximation may be negative. To find such an expansion, we aim to minimize the Kullback-Leibler divergence from π to $\pi_{\mathbf{w}}$. Recall that since the SDE is not fixed *a priori*, we can also optimize over these parameters. We have

$$\min_{\mu, \mathbf{Q}, \Lambda, \mathbf{w}} \mathrm{D}_{\mathrm{KL}}(\pi_{\mathbf{w}} \| \pi) \tag{4.47}$$

where $\mathbf{B} = \mathbf{Q}\Lambda\mathbf{Q}^*$.

This formulation allows us to recast the optimization over $\mu, \mathbf{Q}$, and Λ in terms of a linear transport map. Define function $T(z) = \mu + \mathbf{Q}\Lambda z$. Let $\{e_i\}$ denote the

canonical basis in $\mathbb{R}^d$. Then observe that the pullback of $\pi_{\mathbf{w}}$ under T is

$$\begin{aligned}
\bar{\pi}_{\mathbf{w}} := T^{\#}\pi_{\mathbf{w}} &= \left[1 + \sum_{\mathbf{n}} \mathbf{w}_{\mathbf{n}} \phi_{\mathbf{n}}(\mu + \mathbf{Q}\Lambda z)\right] \eta_\infty(\mu + \mathbf{Q}\Lambda z) |\det \mathbf{Q}\Lambda| \\
&= \left[1 + \sum_{\mathbf{n}} \mathbf{w}_{\mathbf{n}} \prod_{i=1}^{d} \mathrm{He}_{n_i}\left(\frac{\langle \mathbf{Q}\Lambda z, q_i\rangle}{\lambda_i}\right)\right] \frac{\exp\left(-\frac{1}{2}(\mathbf{Q}\Lambda z)^*(\mathbf{B}\mathbf{B}^*)^{-1}(\mathbf{Q}\Lambda z)\right)}{(2\pi)^{d/2}|\det \mathbf{B}\mathbf{B}^*|^{1/2}} |\det \mathbf{Q}\Lambda| \\
&= \left[1 + \sum_{\mathbf{n}} \mathbf{w}_{\mathbf{n}} \prod_{i=1}^{d} \mathrm{He}_{n_i}\left(\frac{\langle z, \Lambda e_i\rangle}{\lambda_i}\right)\right] \frac{\exp\left(-\frac{1}{2} z^* \Lambda \mathbf{Q}^* \mathbf{Q} \Lambda^{-2} \mathbf{Q}^* \mathbf{Q} \Lambda z\right)}{(2\pi)^{d/2}|\det \mathbf{Q}\Lambda^2\mathbf{Q}^*|^{\mathbf{1/2}}} |\det \mathbf{Q}\Lambda| \\
&= \left[1 + \sum_{\mathbf{n}} \mathbf{w}_{\mathbf{n}} \prod_{i=1}^{d} \mathrm{He}_{n_i}(z_i)\right] \frac{\exp\left(-\frac{1}{2} z^* z\right)}{(2\pi)^{d/2}} \\
&= \left[1 + \sum_{\mathbf{n}} \mathbf{w}_{\mathbf{n}} \bar{\phi}_{\mathbf{n}}(z)\right] \bar{\eta}_\infty(z).
\end{aligned}$$

The eigenfunctions $\bar{\phi}_{\mathbf{n}}$ are the tensorized probabilists' Hermite polynomials, and the invariant density $\bar{\eta}_{\mathbf{n}}$ is a standard normal. These correspond to the linear SDE with identity drift and diffusion matrix. With this, we arrive at the optimization problem

$$\min_{T,\mathbf{w}} \mathrm{D}_{\mathrm{KL}}(T_{\#}\bar{\pi}_{\mathbf{w}}, \pi) = \min_{T,\mathbf{w}} \mathrm{D}_{\mathrm{KL}}\left(\bar{\pi}_{\mathbf{w}}, T^{\#}\pi\right). \tag{4.48}$$

Note that the map T is not uniquely determined. A computationally convenient choice of parametrization is to only consider transport maps that are lower triangular. That is, maps of the form $\tilde{T}(z) = \mathbf{L}z + \mu$ where $\mathbf{L}$ is a lower triangular matrix. We can then derive $\mathbf{Q}$ and Λ from $\mathbf{L}$. That is, if $T(z) = \mathbf{Q}\Lambda z + \mu$ will have the same pushforward measure as $\tilde{T}$ if $\mathbf{Q}$ and Λ are derived from the eigenvalue decomposition of $\mathbf{L}\mathbf{L}^*$.

We consider optimization over these two parameters separately. First note that we have

$$\mathrm{D}_{\mathrm{KL}}\left(\bar{\pi}_{\mathbf{w}} \| T^{\#}\pi\right) = \mathbb{E}_{\bar{\pi}_{\mathbf{w}}}\left[\log \bar{\pi}_{\mathbf{w}} - \log T^{\#}\pi\right].$$

When considering optimizing over T alone, we have

$$\min_{T} \mathbb{E}_{\bar{\pi}_{\mathbf{w}}}\left[-\log T^{\#}\pi\right] \tag{4.49}$$

while optimizing over $\mathbf{w}$ requires taking the entire objective into account. The first problem can be solve by existing linear transport map algorithms. For the second problem, we need to compute the gradient. One can derive

$$\frac{\partial}{\partial w_{\mathbf{n}}} \mathrm{D}_{\mathrm{KL}}(\bar{\pi}_{\mathbf{w}} \| T^{\#}\pi) = \int \bar{\phi}_{\mathbf{n}}(x)\bar{\eta}_{\infty}(x)\big[\log \bar{\pi}_{\mathbf{w}}(x) - \log T^{\#}\pi(x)\big]\mathrm{d}x$$

From a computational perspective, we may wish to do importance sampling when computing the gradient since we will likely have samples of $\bar{\pi}_{\mathbf{w}}$ in the course of solving the optimization problem. The exact strategy to use to solve this problem is yet to be determined. Some ideas include alternating minimization, a greedy approach, or a more direct gradient descent approach.

The second question to answer is to figure out how to sample from $\pi_{\mathbf{w}}$. If we can procure samples from $\bar{\pi}_{\mathbf{w}}$, then samples of $\pi_{\mathbf{w}}$ can be found by simply mapping them through the transport map T. Thus, the task at hand is how to obtain from $\bar{\pi}_{\mathbf{w}}$ so that the objective function and its gradient can be evaluated efficiently. These details will be further explored in future iterations of this algorithm.

Chapter 5

Geometry-informed irreversible perturbations for accelerated convergence of Langevin dynamics

5.1 Introduction

In this chapter we consider a different type of stochastic dynamical system that samples from target distributions of interest. We develop new ways for accelerating the convergence of Langevin dynamics (LD) and improve the performance of Langevin-based samplers more broadly. Langevin dynamics-based samplers are often used when one only has access to the target distribution up to a normalizing constant. Langevin dynamics uses the gradient of the log-target density to define the drift term of an SDE whose invariant distribution is the target distribution.

Langevin samplers, however, can often suffer from poor performance when the underlying dynamics converges slowly towards the stationary distribution. It is known that certain perturbations to LD can accelerate convergence to the stationary distribution. In [95] the authors show that suitable reversible and irreversible perturbations to diffusion processes can decrease the spectral gap of the generator, as well as increase the large deviations rate function and decrease the asymptotic variance of the

estimators. Riemannian manifold Langevin dynamics is an example of a reversible perturbation. A simple irreversible sampler adds a term to the drift that is equal to a constant skew-symmetric matrix multiplied by the gradient of the log-target density.

In this chapter we present a *state-dependent irreversible* perturbation of Riemannian manifold Langevin dynamics that is informed by the *geometry* of the manifold. This departs from existing literature, as the vector field of the resulting perturbation is *not* orthogonal to the original drift term. This geometry-informed irreversible perturbation accelerates convergence and, if desired, can be used in combination with the the stochastic gradient Langevin algorithm algorithm to exploit the computational savings of a stochastic gradient.

Traditional sampling methods for Bayesian inference are often intractable for extremely large datasets. While Langevin dynamics-based sampling methods only require access to the unnormalized posterior density, they need many evaluations of this unnormalized density and its gradient. When the dataset is extremely large, each evaluation of the density may be computationally intractable, as it requires the evaluation of the likelihood over the entire dataset. In the past decade the *stochastic gradient* Langevin dynamics (SGLD) has been introduced and analyzed [114, 132] to address the problem posed by large datasets. Rather than evaluating the likelihood over the entire dataset, SGLD subsamples a portion of the data (either with or without replacement) and uses the likelihood evaluated at the sampled data to estimate the true likelihood. The resulting chain can then be used to estimate ergodic averages.

We demonstrate GiIrr on a variety of examples: a simple anisotropic Gaussian target, a posterior on the mean and variance parameters of a normal distribution, Bayesian logistic regression, and Bayesian independent component analysis (ICA). Generally, we observe that the geometry-informed irreversible perturbation improves the convergence rate of LD compared to a standard irreversible perturbation. The improvement tends to be more pronounced as the target distribution deviates from Gaussianity. Our numerical studies also show that introducing irreversibility can reduce the MSE of the resulting long-term average estimator, mainly by reducing variance. In many cases this reduction can be significant, e.g., 1–2 orders of magnitude.

One must, however, also take the effects of discretization into account. In the continuous-time setting, it is known theoretically that irreversible perturbations can at worst only leave the spectral gap fixed. In borderline cases, though—i.e., in cases where the continuous-time theoretical improvement is nearly zero—after accounting for discretization, stiffness can actually cause the resulting estimator to perform worse than if no irreversibility were applied at all. Indeed, we will describe in Appendix B an illustrative Gaussian example in which the standard Langevin algorithm performs better than the algorithm with the standard irreversible perturbation. That is, an example in which additional irreversibility leads to increased bias and variance of the long term average estimator (see Remark 4 for a theoretical explanation). Along similar lines, the idea of applying irreversible perturbations to SGLD has recently been studied in the context of non-convex stochastic optimization [60]. The authors also note that while irreversibility increases the rate of convergence, it increases the discretization error and amplifies the variance of the gradient, compared to a non-perturbed system with the same step size; see also [15] for a related discussion on the relation of SGLD to SGD and convergence properties. This reflects the increased stiffness of irreversible SGLD relative to standard SGLD.

The rest of the chapter is organized as follows. In Section 5.2 we review reversible and irreversible perturbations of the overdamped Langevin dynamics that may improve the efficiency of sampling from equilibrium. Then, in Section 5.2.3, we present our new geometry-informed irreversible perturbation. In Section 5.3 we present simulation studies that demonstrate the good performance of this geometric perturbation, relative to a variety of other standard reversible and irreversible choices. In several of these examples, we also demonstrate the use of stochastic gradients. Section 5.4 summarizes our results and outlines directions for future work. Appendix B details the simple Gaussian example showing that in "borderline" cases—i.e., when continuous-time analysis does not predict improvements from irreversible perturbations—the stiffness created by an irreversible perturbation can, after discretization, lead to poorer performance than the unperturbed case.

5.2 Improving the performance of Langevin samplers

We begin by recalling some relevant background on Langevin samplers, Riemannian manifold Langevin dynamics, perturbations of Langevin dynamics, and the stochastic gradient Langevin dynamics algorithm. Let $f(\theta)$ be a test function on state space $E \subset \mathbb{R}^d$ and let $\pi(\theta)$ be some unnormalized target density on E. In our experiments, $\pi(\theta)$ arises as a posterior density of the form $\pi(\theta) \propto L(\theta; X)\pi_0(\theta)$, where $L(\theta; X)$ is the likelihood model, X are the data, and $\pi_0(\theta)$ is the prior density. Define $\{\theta(t)\}$ as a Langevin process that has invariant density $\pi(\theta)$:

$$\mathrm{d}\theta(t) = \beta \nabla \log \pi(\theta(t))\mathrm{d}t + \sqrt{2\beta}\mathrm{d}W(t), \tag{5.1}$$

where $\beta > 0$ denotes the temperature, $W(t)$ is a standard Brownian motion in $\mathbb{R}^d$, and the initial condition may be arbitrary. Assuming π is twice-continuously differentiable, by ergodicity, we may compute expectations with respect to the posterior by the long term average of $f(\theta)$ over a single trajectory:

$$\mathbb{E}_\pi[f(\theta)] = \int_E f(\theta)\pi(\theta)\mathrm{d}\theta = \lim_{t\to\infty} \frac{1}{T} \int_0^T f(\theta(t))\mathrm{d}t. \tag{5.2}$$

For practical computations, we must approximate (5.2) by discretizing the Langevin dynamics and choosing a large but finite T. Applying the Euler-Maruyama method to (5.1) with step size h yields the following recurrence relation,

$$\theta_{k+1} = \theta_k + h\beta \nabla \log \pi(\theta_t)\mathrm{d}t + \sqrt{2\beta h}\xi_{k+1} \tag{5.3}$$

where ξ_k are independent standard normal random variables. The total number of steps is equal to $K = T/h$. The resulting estimator for (5.2) is

$$\mathbb{E}_\pi[f(\theta)] \approx \frac{1}{K} \sum_{k=0}^{K-1} f(\theta_k). \tag{5.4}$$

This estimator is the *unadjusted Langevin algorithm* (ULA), which has found renewed interest in the context of high-dimensional machine learning problems [44]. Discretization and truncation, however, introduce bias into the estimator. Moreover, there are noted examples in which the continuous-time process and the discretized version do not have the same invariant distribution no matter the choice of the fixed, but nonzero, discretization step h; see [50] for a related discussion. Certain Markov chain Monte Carlo (MCMC) methods such as MALA circumvent these issues by using the dynamics to propose new points, but accepting or rejecting them according to some rule so that the resulting discrete-time Markov chain has the target distribution as its invariant distribution [53, 100].

Many different SDEs can have the same invariant distribution. Therefore, there has been much study into how the standard Langevin dynamics of some target distribution can be altered to increase its rate of convergence. Some examples of this can be found in the work of [61, 97] and others. The standard Langevin dynamics is a reversible Markov process, meaning that the process satisfies detailed balance. The work of [97] studies, in general terms, how reversible and irreversible perturbations to reversible processes decrease the spectral gap, increase the large deviations rate function, and decrease the asymptotic variance. Yet how to *choose* such perturbations to most efficiently accelerate convergence is yet to be thoroughly studied in settings beyond linear diffusion processes [72]. Also, with the exception of a few examples—see for instance [39, 77]—these perturbations have mainly been studied in the continuous-time setting.

5.2.1 Reversible perturbations and Riemannian manifold Langevin dynamics

We only review relevant aspects of reversible perturbations and RMLD in this section. For a detailed review of RMLD and its related Monte Carlo methods, we refer the reader to [53, 76, 134]. Let $\mathbf{B}(\theta)$ be a $d \times d$ symmetric positive definite matrix. A

reversible perturbation on LD (5.1) is an SDE with multiplicative noise:

$$\mathrm{d}\theta(t) = \beta[\mathbf{B}(\theta)\nabla \log \pi(\theta(t)) + \nabla \cdot \mathbf{B}(\theta)]\,\mathrm{d}t + \sqrt{2\beta\mathbf{B}(\theta)}\mathrm{d}W(t). \tag{5.5}$$

Here, the i-th component of $\nabla \cdot \mathbf{B}(\theta)$ is $\sum_{j=1}^{d} \partial_{\theta_j}\mathbf{B}_{ij}(\theta)$. This is equivalent to Langevin dynamics defined on a Riemannian manifold, where the metric is $\mathbf{G}(\theta) = \mathbf{B}(\theta)^{-1}$ [134]. A straightforward calculation shows that (5.5) with $\mathbf{B}(\theta)$ being any symmetric positive-definite matrix admits the same invariant distribution, π. The improved rate of convergence depends on the choice of the underlying metric. The work of [53] argues that choosing the expected Fisher information matrix plus the Hessian of the log-prior to be the metric improves the performance of the resulting manifold MALA method. Meanwhile, [97] shows that under certain regularity conditions, if $\mathbf{B}(\theta)$ is chosen such that $\mathbf{B}(\theta) - \mathbf{I}$ is positive definite, then the resulting estimator is expected to have improved performance in terms of the asymptotic variance, the spectral gap, and the large deviations rate function.

5.2.2 Irreversible perturbations

Consider the following Langevin dynamics

$$\mathrm{d}\theta(t) = [\beta\nabla \log \pi(\theta(t)) + \gamma(\theta(t))]\mathrm{d}t + \sqrt{2\beta}\mathrm{d}W(t). \tag{5.6}$$

When $\gamma(\theta) \equiv 0$, the process is reversible and has $\pi(\theta)$ as its invariant distribution. If $\gamma \neq 0$, then the resulting process will, in general, be time-irreversible unless $\gamma(\theta)$ can be written as a multiple of $\nabla \log \pi(\theta)$; see for example [90]. However, an irreversible perturbation can still preserve the invariant distribution of the system. By considering the Fokker-Planck equation, one can show that if $\gamma(\theta)$ is chosen such that $\nabla \cdot (\gamma\pi) = 0$, then π will still be the invariant distribution. A frequently used choice in the literature is $\gamma(\theta) = \mathbf{J}\nabla \log \pi(\theta)$, where $\mathbf{J}$ is a constant skew-symmetric matrix, i.e., $\mathbf{J} = -\mathbf{J}^T$. The computational advantage of this choice is clear since only one additional matrix-vector multiply is needed to implement this choice. The optimal choice of irreversible

perturbation to linear systems that accelerated convergence fastest was completely analyzed in [72].

The advantages of using irreversible perturbations is widely noted. The main result of [61] is that under certain conditions, the spectral gap, i.e., the difference between the leading two eigenvalues of the generator of the Markov semigroup, increases when $\gamma \neq 0$. In [95, 96, 97], the large deviations rate function is introduced as a measure of performance in the context of sampling from the equilibrium, and upon connecting it to the asymptotic variance of the long term average estimator, it is proven that adding an appropriate perturbation γ not only increases the large deviations rate function but also decreases the asymptotic variance of the estimator. The use of irreversible proposals in the MALA was studied in [86].

5.2.3 Irreversible perturbations for RMLD

In this section, we will introduce our novel geometry-informed irreversible perturbation to Langevin dynamics. Suppose that we are given a diffusion process as in (5.5), and we want to study how to choose an irreversible perturbation that leaves the invariant distribution fixed. Indeed, our previous choice of irreversible perturbation remains valid for this system, that is, adding $\gamma(\theta) = \mathbf{J}\nabla \log \pi(\theta)$ for a constant skew-symmetric matrix $\mathbf{J}$ to the drift term of (5.5) will preserve the invariant density. This choice yields the following SDE:

$$\mathrm{d}\theta(t) = [(\beta \mathbf{B}(\theta(t)) + \mathbf{J})\nabla \log \pi(\theta(t)) + \beta \nabla \cdot \mathbf{B}(\theta(t))]\mathrm{d}t + \sqrt{2\beta \mathbf{B}(\theta)}\mathrm{d}W(t) \tag{5.7}$$

We refer to this system as Riemannian manifold Langevin with an additive irreversible perturbation (`RMIrr`). This choice, however, does not take into account the relevant features that the reversible perturbation may provide when constructing an irreversible perturbation.

The reversible perturbation leads to a positive definite matrix (a metric, in the terminology of Riemannian geometry) that is state-dependent. In contrast, the skew-symmetric matrix $\mathbf{J}$ is fixed in the irreversible perturbation. The skew-symmetric

matrix need not be constant, however, as an irreversible perturbation $\gamma(\theta)$ only needs to satisfy $\nabla \cdot (\gamma(\theta)\pi(\theta)) = 0$. In fact, if $\gamma(\theta) = \mathbf{C}(\theta)\nabla \log \pi(\theta) + \nabla \cdot \mathbf{C}(\theta)$ for $\mathbf{C}(\theta) = -\mathbf{C}(\theta)^T$, then this irreversible perturbation will also leave the invariant density intact. Noting that $\mathbf{C}_{ii}(\theta) = 0$ and that $\mathbf{C}_{ij} = -\mathbf{C}_{ji}$, observe that

$$\begin{aligned}
\nabla \cdot (\gamma(\theta)\pi(\theta)) &= \nabla \cdot (\mathbf{C}(\theta)\nabla\pi(\theta) + (\nabla \cdot \mathbf{C}(\theta))\pi(\theta)) \\
&= \sum_{i,j=1}^{d} \frac{\partial \mathbf{C}_{ij}(\theta)}{\partial \theta_i}\frac{\partial \pi(\theta)}{\partial \theta_j} + \mathbf{C}_{ij}(\theta)\frac{\partial^2 \pi(\theta)}{\partial \theta_i \partial \theta_j} + \frac{\partial^2 \mathbf{C}_{ij}(\theta)}{\partial \theta_i \partial \theta_j}\pi(\theta) + \frac{\partial \mathbf{C}_{ij}(\theta)}{\partial \theta_j}\frac{\partial \pi(\theta)}{\partial \theta_i} \\
&= \sum_{i>j,i=1}^{d} \frac{\partial \mathbf{C}_{ij}(\theta)}{\partial \theta_i}\frac{\partial \pi(\theta)}{\partial \theta_j} + \mathbf{C}_{ij}(\theta)\frac{\partial^2 \pi(\theta)}{\partial \theta_i \partial \theta_j} + \frac{\partial^2 \mathbf{C}_{ij}(\theta)}{\partial \theta_i \partial \theta_j}\pi(\theta) + \frac{\partial \mathbf{C}_{ij}(\theta)}{\partial \theta_j}\frac{\partial \pi(\theta)}{\partial \theta_j} \\
&\quad + \frac{\partial \mathbf{C}_{ji}(\theta)}{\partial \theta_j}\frac{\partial \pi(\theta)}{\partial \theta_i} + \mathbf{C}_{ji}(\theta)\frac{\partial^2 \pi(\theta)}{\partial \theta_j \partial \theta_i} + \frac{\partial^2 \mathbf{C}_{ji}(\theta)}{\partial \theta_j \partial \theta_i}\pi(\theta) + \frac{\partial \mathbf{C}_{ji}(\theta)}{\partial \theta_i}\frac{\partial \pi(\theta)}{\partial \theta_i} \\
&= \sum_{i>j,i=1}^{d} \frac{\partial \mathbf{C}_{ij}(\theta)}{\partial \theta_i}\frac{\partial \pi(\theta)}{\partial \theta_j} + \mathbf{C}_{ij}(\theta)\frac{\partial^2 \pi(\theta)}{\partial \theta_i \partial \theta_j} + \frac{\partial^2 \mathbf{C}_{ij}(\theta)}{\partial \theta_i \partial \theta_j}\pi(\theta) + \frac{\partial \mathbf{C}_{ij}(\theta)}{\partial \theta_j}\frac{\partial \pi(\theta)}{\partial \theta_i} \\
&\quad - \frac{\partial \mathbf{C}_{ij}(\theta)}{\partial \theta_j}\frac{\partial \pi(\theta)}{\partial \theta_i} - \mathbf{C}_{ij}(\theta)\frac{\partial^2 \pi(\theta)}{\partial \theta_j \partial \theta_i} - \frac{\partial^2 \mathbf{C}_{ij}(\theta)}{\partial \theta_j \partial \theta_i}\pi(\theta) - \frac{\partial \mathbf{C}_{ij}(\theta)}{\partial \theta_i}\frac{\partial \pi(\theta)}{\partial \theta_j} \\
&= 0.
\end{aligned}$$

We seek an irreversible perturbation that takes the reversible perturbation into account, with the possibility that $\mathbf{C}(\theta)$ is not a constant matrix, and investigate if it leads to any performance improvements of the long term average estimator. Note that in the literature, the above condition $\nabla \cdot (\gamma\pi) = 0$ is typically rewritten into the following sufficient conditions: $\nabla \cdot \gamma(\theta) = 0$ and $\gamma(\theta) \cdot \nabla\pi(\theta) = 0$ [97]. One can check, however, that when $\mathbf{C}$ is not constant, these conditions are not met, yet $\gamma(\theta)$ is still a valid irreversible perturbation. A simple choice of $\mathbf{C}(\theta)$ that incorporates $\mathbf{B}(\theta)$ is

$$\mathbf{C}(\theta) = \frac{1}{2}\mathbf{J}\mathbf{B}(\theta) + \frac{1}{2}\mathbf{B}(\theta)\mathbf{J}\,, \tag{5.8}$$

where $\mathbf{J}$ is a constant skew-symmetric matrix. The $\frac{1}{2}$ factor is introduced so that if $\mathbf{B}(\theta) = \mathbf{I}$, i.e., if there is no reversible perturbation, then this perturbation reverts to

the standard irreversible perturbation (`Irr`). We arrive at the following system:

$$\begin{aligned} \mathrm{d}\theta(t) =& [(\beta \mathbf{B}(\theta(t)) + \mathbf{C}(\theta(t)))\nabla \log \pi(\theta(t)) + \nabla \cdot (\beta \mathbf{B}(\theta(t)) + \mathbf{C}(\theta(t))]\mathrm{d}t \\ &+ \sqrt{2\beta \mathbf{B}(\theta(t))}\mathrm{d}W(t). \end{aligned} \tag{5.9}$$

We call this choice of perturbation the *geometry-informed irreversible perturbation* (`GiIrr`). Indeed, while there are infinitely many valid choices for $\mathbf{C}(\theta)$, we will investigate the choice in (5.8) in the numerical examples. Since we will have already explicitly constructed $\mathbf{B}(\theta)$ and $\mathbf{J}$ for the other systems, the additional computational cost of computing their product will be marginal. Furthermore, as mentioned earlier, this choice reduces to `Irr` when $\mathbf{B}(\theta) = \mathbf{I}$.

One may wonder when does `GiIrr` result in improved performance over standard irreversible perturbations such as in (5.7)? Based on the numerical results and intuition, we will argue that `GiIrr` results in better performance if the underlying reversible perturbation already improves the sampling. As we mentioned earlier, the choice of `GiIrr` that is made in this chapter is not unique, and a further investigation of its theoretical properties is left for future work; see also the discussion in the Section 5.4. The goal is to present this new class of irreversible perturbations and investigate it numerically in a number of representative computational studies.

5.2.4 Stochastic gradient Langevin dynamics

In certain Bayesian inference problems, the data are conditionally independent of each other given the parameter value. Therefore, the likelihood model can often be factorized and the posterior density can be written as follows:

$$\pi(\theta) \propto \pi_0(\theta) \prod_{i=1}^{N} \pi_i(X_i|\theta) \tag{5.10}$$

where $\pi(X_i|\theta)$ is the likelihood function for data point X_i. When the dataset is extremely large, i.e., when $N \gg 1$, however, ULA becomes exceedingly expensive as it requires repeatedly evaluating the likelihood over the entire dataset for each step of

	$b(\theta)$	$\sigma(\theta)$
`LD`	$\beta\nabla \log \pi(\theta)$	$\sqrt{2\beta}\mathbf{I}$
`RM`	$\beta\mathbf{B}(\theta)\nabla \log \pi(\theta)) + \beta\nabla \cdot \mathbf{B}(\theta)$	$\sqrt{2\beta\mathbf{B}(\theta)}$
`Irr`	$(\beta\mathbf{I} + \mathbf{J})\nabla \log \pi(\theta)$	$\sqrt{2\beta}\mathbf{I}$
`RMIrr`	$(\beta\mathbf{B}(\theta) + \mathbf{J})\nabla \log \pi(\theta) + \beta\nabla \cdot \mathbf{B}(\theta)$	$\sqrt{2\beta\mathbf{B}(\theta)}$
`GiIrr`	$(\beta\mathbf{B}(\theta) + \frac{1}{2}\mathbf{J}\mathbf{B}(\theta) + \frac{1}{2}\mathbf{B}(\theta)\mathbf{J})\nabla \log \pi(\theta)$ $+\nabla \cdot (\beta\mathbf{B}(\theta) + \frac{1}{2}\mathbf{J}\mathbf{B}(\theta) + \frac{1}{2}\mathbf{B}(\theta)\mathbf{J})$	$\sqrt{2\beta\mathbf{B}(\theta)}$

Table 5.1: Summary of the five SDEs that share the same invariant density $\pi(\theta)$. Stochastic gradients can be considered instead of the deterministic gradients. All systems are of the form $\mathrm{d}\theta_t = b(\theta_t)\mathrm{d}t + \sigma(\theta_t)\mathrm{d}W_t$. The term β denotes the temperature.

the trajectory. To mitigate this challenge, the stochastic gradient Langevin dynamics was presented to reduce the computational cost of evaluating the posterior density by only evaluating the likelihood over *subsets* of the data at each step. The true likelihood is estimated based on the likelihood function evaluated at the subsampled data [132]. Specifically, the gradient is estimated using a stochastic gradient

$$\nabla \log \pi(\theta|X) \approx \nabla \widehat{\log \pi(\theta}|X) = \log \pi_0(\theta) + \frac{N}{n}\sum_{i=1}^{n} \log \pi(X_{\tau_i}|\theta) \tag{5.11}$$

where τ is a random subset of $\{1, \ldots, N\}$ of size n drawn with or without replacement. Depending on the choice of n, this approach cuts down on the computational costs dramatically with some additional variance incurred by the random subsampling of the data. The original version of this algorithm made the step size variable, approaching zero as the number of steps taken K became large. SGLD applied with a variable and shrinking step size was proven to be consistent: that is, the invariant distribution of the discretized system is equivalent to that of the continuous system [114]. Having a decreasing step size counteracts the cost savings provided by computing the stochastic gradient, and therefore a version where the step size is fixed was presented in [128], where theoretical characterizations of the asymptotic and finite-time bias and variance are also developed. In most of our numerical results, we use stochastic gradient version of the Langevin algorithm with fixed step size to demonstrate that SGLD can be used together with irreversible perturbations.

5.3 Numerical examples

In the following examples, we always apply the stochastic gradient version of each Langevin system unless otherwise stated. We fix $\beta = 1/2$ for all examples. The efficacy of the `GiIrr` perturbation does not change whether or not the stochastic gradient is used. We illustrate this explicitly in Section 5.3.3, where we report the results of all perturbations both with and without the stochastic gradient, for comparison.

5.3.1 Linear Gaussian example

Suppose we have data $\{X_i\}_{i=1}^N \subset \mathbb{R}^d$ generated from a bivariate normal distribution with mean $\theta \in \mathbb{R}^d$ and known precision matrix $\mathbf{\Gamma}_X \in \mathbb{R}^{d\times d}$. From the data, we infer the value of θ. Endow θ with a normal prior with mean zero and precision $\mathbf{\Gamma}_\theta \in \mathbb{R}^{d\times d}$. Then the posterior distribution is Gaussian with mean and precision

$$\mu_p = (\mathbf{\Gamma}_\theta + N\mathbf{\Gamma}_X)^{-1}\mathbf{\Gamma}_X \sum_{i=1}^N X_i \quad \text{and} \quad \mathbf{\Gamma}_p = (\mathbf{\Gamma}_\theta + N\mathbf{\Gamma}_X), \tag{5.12}$$

respectively. The Euler-Maruyama discretization with constant step size h applied to the corresponding Langevin dynamics is

$$\theta_{k+1} = \big(\mathbf{I} - \bar{\mathbf{A}}h\big)\theta_k + \bar{\mathbf{D}}_k h + \sqrt{h}\xi_k \tag{5.13}$$

where

$$\bar{\mathbf{A}} = \frac{1}{2}(\mathbf{\Gamma}_\theta + N\mathbf{\Gamma}_X), \quad \bar{\mathbf{D}}_k = \frac{1}{2}\mathbf{\Gamma}_X \sum_{i=1}^N X_i, \quad \xi_k \sim \mathcal{N}(0, \mathbf{I}).$$

Using stochastic gradients yields the same recurrence above except with

$$\bar{\mathbf{D}}_k = \frac{1}{2}\mathbf{\Gamma}_X \frac{N}{n} \sum_{i=1}^n X_{\tau_i^k} \tag{5.14}$$

where $n \leq N$ and $\tau_i^k \in \{1, \ldots, N\}$ is randomly sampled (with or without replacement) [132]. Expectations with respect to the posterior are approximated by an long term

average of the observable over the course of a trajectory. It has been shown that despite subsampling the data at each step in the dynamics, this estimator has comparable performance as the estimator produced by the regular Langevin dynamics with the full likelihood or MALA [128, 132].

Now, we consider the case where the dynamics are perturbed by an irreversible term that preserves the invariant distribution of the dynamics. We demonstrate that this leads to a lower MSE than standard SGLD or Langevin dynamics. In this case, we replace $\bar{\mathbf{A}}$ and $\bar{\mathbf{D}}_k$ with $\mathbf{A}$ and $\mathbf{D}_k$, which are

$$\mathbf{A} = \frac{1}{2}(\mathbf{I} + \mathbf{J})(\boldsymbol{\Gamma}_\theta + N\boldsymbol{\Gamma}_X), \quad \mathbf{D}_k = \frac{1}{2}(\mathbf{I} + \mathbf{J})\boldsymbol{\Gamma}_X \frac{N}{n} \sum_{i=1}^{n} X_{\tau_i^k}. \tag{5.15}$$

and $\mathbf{J}$ is a skew-symmetric matrix.

For the numerical experiments, we choose $d = 3$, $N = 10$, where the mini-batches are of size $n = 2$. We have $\boldsymbol{\Gamma}_X = 0.25\mathbf{I}$, $\boldsymbol{\Gamma}_\theta$ is a precision matrix with eigenvalues $0.2, 0.01, 0.05$ and eigenvectors that are randomly generated, and $h = 0.005$. Note that these matrices were chosen so that the resulting reversible perturbation has eigenvalues greater than one. To construct the perturbations, we choose $\mathbf{B} = \boldsymbol{\Gamma}_p^{-1}$ and $\mathbf{J}$ to be

$$\mathbf{J} = \delta \begin{bmatrix} 0 & 1 & 1 \\ -1 & 0 & 1 \\ -1 & -1 & 0 \end{bmatrix} \tag{5.16}$$

for $\delta \in \mathbb{R}$. We consider the five different SDE systems presented in Table 5.1 and investigate how the MSE, bias, and variance differs for each case. For this example, since a constant metric is used, the geometry-informed irreversible perturbation simply produces a different constant skew-symmetric matrix than the other irreversible perturbations. Each system is simulated for $K = 10^5$ steps with step size $h = 5 \times 10^{-3}$. In Figure 5-1, we plot the MSE of the running average for each case when the observables are the sums of the first and second moments. To compute the asymptotic variances of the observables we use the batch means method in [4]. After the burn-in period of $T_b = 5$, we evaluate the observable over each chain. Each observable chain

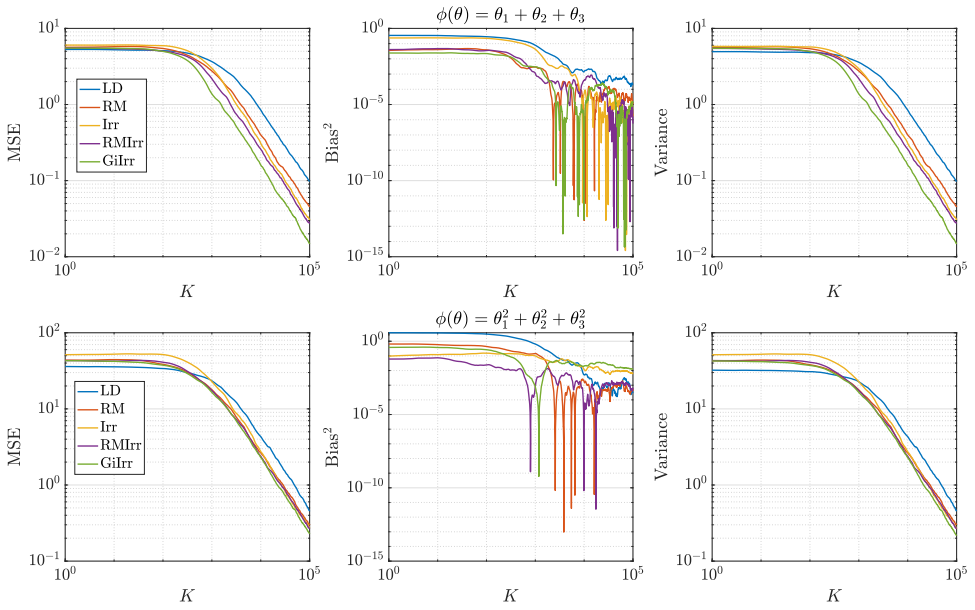


Figure 5-1: MSE of the running average for the first and second moments. Stochastic gradients are used in this example.

is then batched into twenty separate chains, and their means are evaluated. The asymptotic variance is estimated by computing the empirical variance of those means and then multiplying by the length of each of the subsampled trajectories. In Table 5.2, we report the asymptotic variance of the estimator for each system.[1] We see that irreversible perturbations definitely improve the performance of the estimators, although the improvement provided by the geometry-informed irreversible perturbation seems marginal over `RMIrr` when estimating the second moments.

When the reversible perturbation is chosen such that the drift matrix is exactly the identity (for example, when the matrix is chosen to be the covariance matrix of the posterior), additional irreversibility cannot widen the spectral gap of the system. This fact can be deduced from the results of [72]. The improved performance of the geometry-informed irreversible perturbation is mostly due to the fact that the norm of the corresponding skew-symmetric matrix is greater than that of simple irreversibility. Even though one can scale the skew-symmetric matrix for the other two cases to observe similar performance as geometry-informed irreversibility, `GiIrr` accomplishes

[1]The asymptotic variance reported here is σ^2 where $\mathrm{Var}(Y(t)) \sim \sigma^2/t$, $Y(t) = \frac{1}{t}\int_0^t \phi(\theta_t)dt$, and ϕ is an observable.

that in a more systematic way.

	$\mathbb{E}[\text{AVar}_{\phi_1}]$	$\text{Std}[\text{AVar}_{\phi_1}]$	$\mathbb{E}[\text{AVar}_{\phi_2}]$	$\text{Std}[\text{AVar}_{\phi_2}]$
`LD`	37.75	11.94	209.4	84.38
`RM`	20.09	6.420	132.8	49.21
`Irr`	15.72	5.008	135.4	47.91
`RMIrr`	12.36	3.937	115.9	40.10
`GiIrr`	**7.444**	2.336	**103.7**	36.78

Table 5.2: Asymptotic variance estimates for the linear Gaussian example.

While it is known that irreversible perturbations can, at worst, maintain the same performance as standard Langevin in the continuous-time setting [97], when considering discretization and in borderline cases (i.e., when one does not expect much or any improvement in continuous time), irreversibility may actually harm the performance of the estimator as it introduces additional stiffness into the system without resulting in faster convergence to the invariant density. A detailed exploration of this effect is presented in Appendix B, in which we compute the bias and variance of the long term average estimator for a simple linear Gaussian problem where the posterior precision is a scalar multiple of the identity matrix. As further discussed in Remark 4, in this case, the irreversible perturbation is not expected to lead to improvement in the sampling properties from the equilibrium. Hence, the stiffness induced upon discretization has a more profound impact on the practical performance of the irreversible perturbation.

In the current numerical study, the posterior precision is diagonal, but not a scalar multiple of the identity matrix. The eigenvalues of the resulting drift matrix are therefore distinct, and by the theory in [72], irreversible perturbations are able to reduce the spectral gap and result in improved performance. This is in contrast with the example studied in the appendix.

5.3.2 Parameters of a normal distribution

This example is identical to that used in [53, Section 5] to demonstrate the performance of RMLD. Given a dataset of $\mathbb{R}$-valued data $\mathbf{X} = \{X_i\}_{i=1}^N \sim \mathcal{N}(\mu, \sigma^2)$, we infer the parameters μ, σ. To be clear, in this example the state is $\theta = [\mu, \sigma]^T$. The prior on μ, σ is chosen to be flat (and, therefore, improper). The log-posterior is

$$\log p(\mu, \sigma|\mathbf{X}) = \frac{N}{2}\log 2\pi - N\log\sigma - \sum_{i=1}^{N}\frac{(X_i-\mu)^2}{2\sigma^2}. \tag{5.17}$$

The gradient is

$$\nabla \log p(\mu, \sigma|\mathbf{X}) = \begin{bmatrix} m_1(\mu)/\sigma^2 \\ -N/\sigma + m_2(\mu)/\sigma^3 \end{bmatrix} \tag{5.18}$$

where $m_1(\mu) = \sum_{i=1}^N (X_i - \mu)$, and $m_2(\mu) = \sum_{i=1}^N (X_i-\mu)^2$. In [53], the authors propose using the geometry of the manifold defined by the parameter space of the posterior distribution to accelerate the resulting Metropolis-adjusted Langevin algorithm. The authors in [53] suggest using the expected Fisher information matrix to define the Riemannian metric, which in the context of reversible diffusions [97], is equivalent to choosing $\mathbf{B}(\mu, \sigma)$ to be the inverse of the sum of the expected Fisher information matrix and the negative Hessian of the log-prior. Straightforward computations yield

$$\mathbf{B} = \frac{\sigma^2}{N}\begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix}, \quad \sqrt{\mathbf{B}} = \frac{\sigma}{\sqrt{N}}\begin{bmatrix} 1 & 0 \\ 0 & 1/\sqrt{2} \end{bmatrix}, \quad \nabla\cdot\mathbf{B} = \begin{bmatrix} 0 \\ \sigma/N \end{bmatrix}. \tag{5.19}$$

As for the geometry-informed irreversible perturbation, let $\mathbf{J} = \delta\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$, for $\delta \in \mathbb{R}$. Then the relevant quantities are

$$\frac{1}{2}\mathbf{J}\mathbf{B} + \frac{1}{2}\mathbf{B}\mathbf{J} = \frac{3\sigma^2}{4N}\mathbf{J}, \quad \frac{1}{2}\nabla\cdot(\mathbf{J}\mathbf{B} + \mathbf{B}\mathbf{J}) = \frac{3\delta\sigma}{2N}\begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{5.20}$$

In the experiments, we have $N = 30$, $h = 10^{-3}$, $\delta = 2$. and simulate $M = 1000$

independent trajectories of each system up to $T = 1000$ for a total of $K = 10^6$ steps. The data is subsampled at a rate of $n = 6$ per stochastic gradient computation. Each trajectory is allotted a burn-in time of $T_b = 10$. The dataset is generated by drawing samples from a normal distribution with $\mu_{true} = 0$ and $\sigma_{true} = 10$. The observables we study are $\phi_1(\mu, \sigma) = \mu + \sigma$, and $\phi_2(\mu, \sigma) = \mu^2 + \sigma^2$. We plot the MSE, squared bias, and variance of resulting estimators for each observable in Figures 5-3 and 5-4. Moreover, in Table 5.3 we report the asymptotic variance of the estimators of each of the five systems. The main takeaway is that an irreversible perturbation that is adapted to the existing reversible perturbation performs much better than if the irreversible perturbation were applied without regard to the underlying geometry. Notice that the reversible perturbation considered here still improves the performance of the long term average estimator despite the fact that $\mathbf{B} - \mathbf{I}$ is not positive definite on the state space. Indeed, while $\mathbf{B} - \mathbf{I}$ being positive definite is a sufficient condition to obtain improved performance, it is not a necessary one [97]. The reason for the reduced asymptotic variance we observed here is because the reversible perturbation $\mathbf{B}$ has eigenvalues larger than one where the bulk of the posterior distribution lies.

Figure 5-2 show single and mean trajectories of the burn-in period of trajectories from each of the five systems. The plot shows that the geometry-informed irreversible perturbation is able to find the bulk of the distribution sooner than the other systems without incurring additional errors due to stiffness.

To show that the `GiIrr` perturbation is not intimately tied to the stochastic gradient, we also report the results for each system when the gradients are computed exactly in Table 5.4. We see that there is little meaningful difference in the results compared to when stochastic gradients are used.

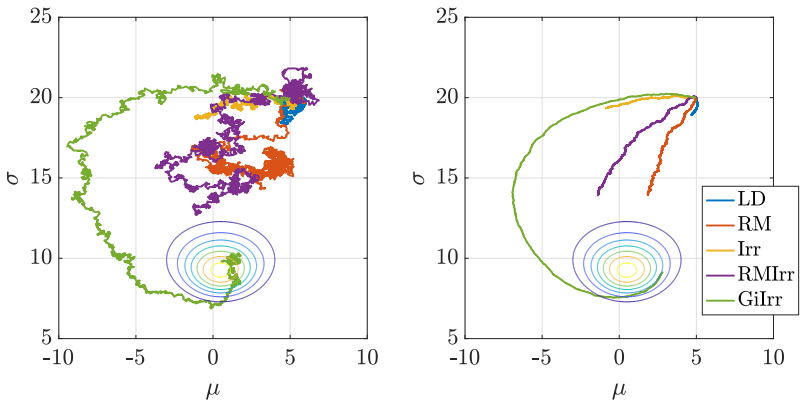


Figure 5-2: Trajectory burn-in: each trajectory is run for $T = 2.5$. Left: single trajectories, right: mean paths. The gradients are computed exactly here.

	$\mathbb{E}[\text{AVar}_{\phi_1}]$	$\text{Std}[\text{AVar}_{\phi_1}]$	$\mathbb{E}[\text{AVar}_{\phi_2}]$	$\text{Std}[\text{AVar}_{\phi_2}]$
`LD`	55.29	21.52	8332	4359
`RM`	20.63	6.019	4034	1378
`Irr`	5.791	2.638	2169	1072
`RMIrr`	6.512	2.226	1729	631.2
`GiIrr`	**1.400**	0.4697	**479.4**	170.8

Table 5.3: Asymptotic variance estimates for the parameters of a normal distribution example. Stochastic gradients are employed.

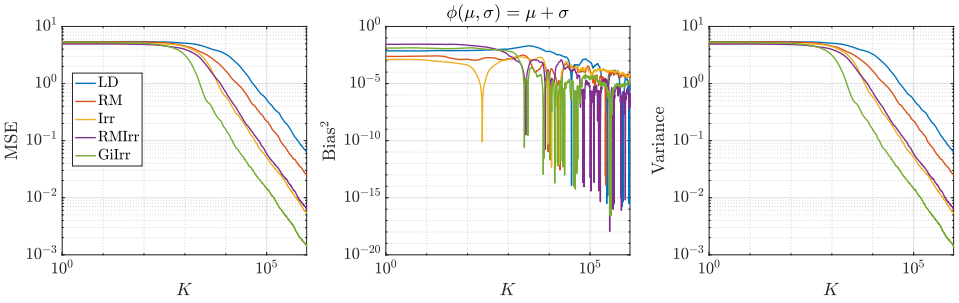


Figure 5-3: Observable: $\phi_1(\mu, \sigma) = \mu + \sigma$, $\delta = 2$. Stochastic gradients are computed.

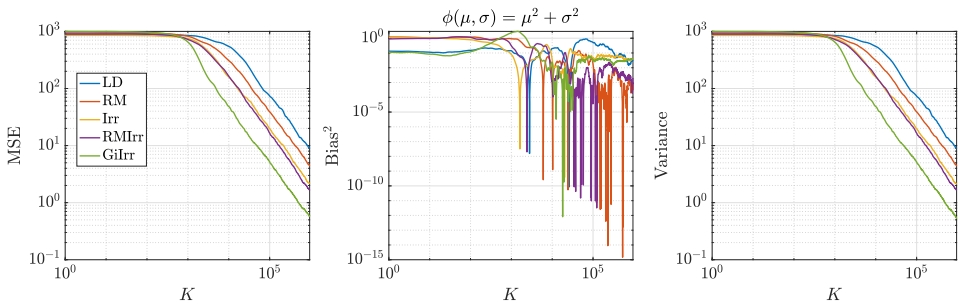


Figure 5-4: Observable: $\phi_2(\mu, \sigma) = \mu^2 + \sigma^2$, $\delta = 2$. Stochastic gradients are computed.

	$\mathbb{E}[\text{AVar}_{\phi_1}]$	$\text{Std}[\text{AVar}_{\phi_1}]$	$\mathbb{E}[\text{AVar}_{\phi_2}]$	$\text{Std}[\text{AVar}_{\phi_2}]$
`LD` (no SG)	48.51	17.53	7339	3707
`RM` (no SG)	20.91	6.445	3855	1406
`Irr` (no SG)	5.658	2.108	2265	1191
`RMIrr` (no SG)	6.276	2.075	1648	565.1
`GiIrr` (no SG)	**1.363**	0.4223	**492.9**	183.8

Table 5.4: Asymptotic variance estimates for the parameters of a normal distribution example. The gradients are computed exactly.

5.3.3 Bayesian logistic regression

Next we consider Bayesian logistic regression. Given data $\{(\mathbf{x}_i, t_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^d$, and $t_i \in \{0, 1\}$, we seek a logistic function, parameterized by weights $\mathbf{w} \in \mathbb{R}^d$, that best fits the data. The weights are obtained in a Bayesian fashion, in which we endow the weights with a prior and seek to characterize its posterior distribution via sampling. Define $\varphi(y)$ to be the logistic function $\varphi(y) = (1 + \exp(-y))^{-1}$. The log-likelihood function is

$$l(\mathbf{w}) = \sum_{i=1}^{N} t_i \mathbf{x}_i^T \mathbf{w} - \sum_{i=1}^{N} \log(1 + \exp(\mathbf{x}_i^T \mathbf{w})). \tag{5.21}$$

The prior for the weights is normally distributed with mean zero and covariance $\alpha^{-1}\mathbf{I}$. The gradient of the log-posterior is

$$\nabla_{\mathbf{w}} \log \pi(\mathbf{w}|\mathbf{X}) = -\alpha\mathbf{w} + \sum_{i=1}^{N} t_i \mathbf{x}_i - \sum_{i=1}^{N} \varphi(\mathbf{x}_i^T \mathbf{w})\mathbf{x}_i. \tag{5.22}$$

This term is used in the drift part of the Langevin dynamics that fully computes the gradient of the log-likelihood at every step. If the data are subsampled as in SGLD, we instead compute

$$\nabla_{\mathbf{w}} \log \tilde{\pi}(\mathbf{w}|\mathbf{X}) = -\alpha\mathbf{w} + \frac{N}{n}\sum_{i=1}^{n} t_{\tau_i} \mathbf{x}_{\tau_i} - \frac{N}{n}\sum_{i=1}^{n} \phi(\mathbf{x}_{\tau_i}^T \mathbf{w})\mathbf{x}_{\tau_i}. \tag{5.23}$$

We use the `german` data set described in [52] for the numerical experiments. In this problem, there are 20 weight parameters to be learned. The training dataset is of size $N = 400$ and we choose to subsample at a rate of $n = 10$ per likelihood computation. The time step we choose is $h = 10^{-4}$ and $K = 4 \times 10^5$ steps. We generate the skew-symmetric matrix by constructing a lower triangular matrix with entries randomly drawn from $\{1, -1\}$ and then subtracting its transpose. The diagonal is then set to zero and the matrix is scaled to have norm one.

As for the Riemannian manifold Langevin dynamics, in [53] the authors use the expected Fisher information matrix plus the negative Hessian of the log-prior as the underlying metric, which in this case is equal to

$$\mathbf{G}(w) = \alpha^{-1}\mathbf{I} + \mathbf{X}\mathbf{\Lambda}(w)\mathbf{X}^T \tag{5.24}$$

where $\mathbf{\Lambda}$ is a diagonal matrix with entries $\mathbf{\Lambda}_{ii}(w) = (1 - \varphi(\mathbf{x}_i^T w))\varphi(\mathbf{x}_i^T w)$ and $\mathbf{x}_i$ is the i-th column of $\mathbf{X}$. The resulting reversible perturbation uses the inverse of $\mathbf{G}(w)$. This perturbation, however, does not lead to accelerated convergence to the invariant measure since the eigenvalues of $\mathbf{G}$ are large. This implies that the eigenvalues of $\mathbf{G}^{-1}$ are less than one and so $\mathbf{G}^{-1}(w) - \mathbf{I}$ is not positive definite, a condition that needs to be satisfied to guarantee accelerated convergence [95]. To alleviate this

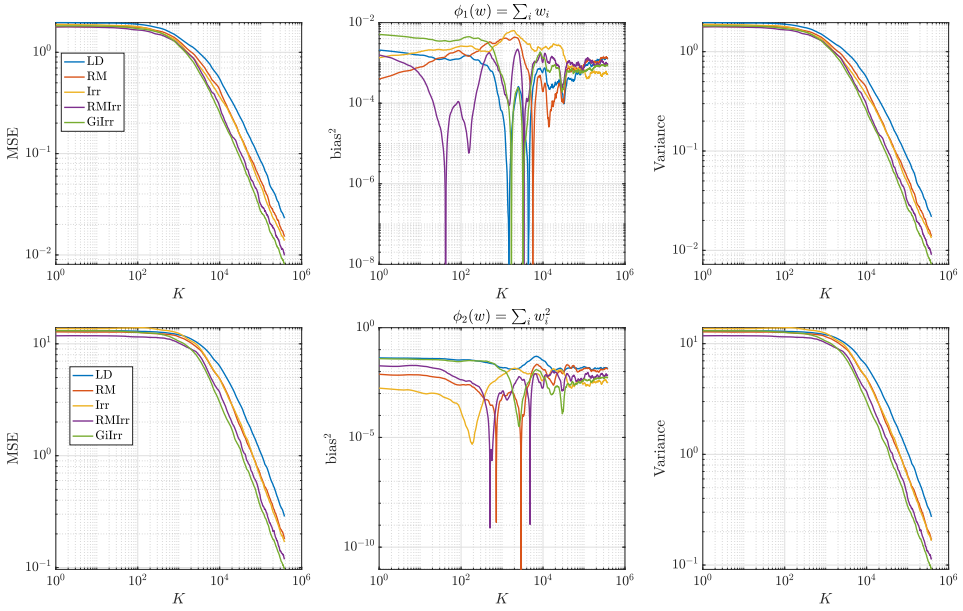


Figure 5-5: Bayesian logistic regression with a variable metric. Here, $d = 20$.

issue, we consider the reversible perturbation $\mathbf{B}(w) = \mathbf{I} + \mathbf{G}^{-1}(w)$. This guarantees $\mathbf{B}(w)$ to be positive definite for all w, but the drawback is that computing the square root of $\mathbf{B}(w)$ requires explicitly computing or at least approximating the inverse of $\mathbf{G}(w)$ repeatedly in the simulation (and not just computing the action of the inverse). This additional computational cost is incurred for all examples that consider a geometry-informed perturbation, both reversible and irreversible. We show the result of this state-dependent perturbation in Figure 5-5 and report the asymptotic variance in Table 5.5. The geometry-informed irreversible perturbation does provide improvement over all other perturbations. We observe that the asymptotic variance is reduced by half over `RM`, with only little additional computational effort. Most of the computational cost of applying `GiIrr` is due to the evaluation of the reversible perturbation. Therefore we emphasize that if one is already applying the reversible perturbation to the Langevin dynamics, the marginal cost of applying the `GiIrr` perturbation is negligible.

	$\mathbb{E}[\text{AVar}_{\phi_1}]$	$\text{Std}[\text{AVar}_{\phi_1}]$	$\mathbb{E}[\text{AVar}_{\phi_2}]$	$\text{Std}[\text{AVar}_{\phi_2}]$
`LD`	1.967	0.9995	23.77	12.52
`RM`	1.328	0.6538	15.35	7.348
`Irr`	1.163	0.5698	14.84	7.738
`RMIrr`	0.8775	0.4228	10.68	5.306
`GiIrr`	**0.7148**	0.3450	**8.798**	4.490

Table 5.5: Asymptotic variance estimates for the Bayesian logistic regression example with a state-dependent metric.

5.3.4 Independent component analysis

Our last example considers the problem of blind signal separation addressed in [132] and [1]. This problem yields a posterior that is strongly non-Gaussian and multi-modal, and we show that `GiIrr` has substantially better sampling performance over standard reversible and irreversible perturbations. Suppose there are m separate unknown independent signals $s^i(t)$ for $i = 1, \ldots, m$ that are mixed by mixing matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$. Suppose we can observe the mixed signals $X(t) = \mathbf{M}s(t)$ for N instances in time. The goal of independent component analysis is to infer a de-mixing matrix $\mathbf{W}$ such that the m signals are recovered up to a nonzero constant and permutation. As such, this problem is generally ill-posed, but is suitable to be considered in a Bayesian context. The ICA literature states that, based on real-world data, it is best to assume a likelihood model with large kurtosis. Following [1, 132], let $p(y_i) = \frac{1}{4}\text{sech}^2\left(\frac{1}{2}y_i\right)$. The prior on the weights $\mathbf{W}_{ij}$ is Gaussian with zero mean and precision λ. The posterior is equal to

$$p(\mathbf{W}|X) \propto |\det \mathbf{W}| \prod_{i=1}^{m} p(\mathbf{w}_i^T \mathbf{x}) \prod_{ij} \mathcal{N}(\mathbf{W}_{ij}; 0, \lambda^{-1}). \tag{5.25}$$

The gradient of the log posterior with respect to the matrix $\mathbf{W}$ is then

$$f(\mathbf{W}) = \nabla_{\mathbf{W}} \log p(\mathbf{W}|X) = \left(N(\mathbf{W}^T)^{-1} - \sum_{n=1}^{N} \tanh\left(\frac{1}{2}\mathbf{y}_n\right)\mathbf{x}_n^T \right) - \lambda \mathbf{W}. \tag{5.26}$$

It is suggested in [1] that the natural gradient should be used instead of the gradient we see here above to account for the information geometry of the problem. Specifically, [1, 114] post-multiply the gradient by $\mathbf{W}^T\mathbf{W}$ and arrive at the so-called natural gradient of the system

$$\mathcal{D}_{\mathbf{W}} := \left(N\mathbf{I} - \sum_{n=1}^{N} \tanh\left(\frac{1}{2}\mathbf{y}_n\right)\mathbf{y}_n^T \right)\mathbf{W} - \lambda\mathbf{W}\mathbf{W}^T\mathbf{W}. \tag{5.27}$$

In the context of RMLD, this is equivalent to perturbing the system with a reversible perturbation with $\mathbf{B}(\mathbf{W}) = \mathbf{W}^T\mathbf{W} \otimes \mathbf{I}$ pre-multipled in front of the vectorized gradient. That is, we have

$$\mathrm{vec}(f(\mathbf{W})\mathbf{W}^T\mathbf{W}) = (\mathbf{W}^T\mathbf{W} \otimes \mathbf{I})\mathrm{vec}f(\mathbf{W}).$$

We construct the `GiIrr` term as follows. To take advantage of the matrix structure of the reversible perturbation, we choose the skew-symmetric matrix such that it acts within the computation of the natural gradient. We choose $\mathbf{J} = (\mathbf{I} \otimes \mathbf{C}_0) + (\mathbf{C}_0 \otimes \mathbf{I})$ where $\mathbf{C}_0$ has the same sign pattern as (5.16) but such that $\mathbf{J}$ has matrix norm equal to 2. Then the geometry-informed irreversible perturbation is

$$\frac{1}{2}\mathbf{B}(\mathbf{W})\mathbf{J} + \frac{1}{2}\mathbf{J}\mathbf{B}(\mathbf{W}) = (\mathbf{W}^T\mathbf{W} \otimes \mathbf{C}_0) + \frac{1}{2}(\mathbf{W}^T\mathbf{W}\mathbf{C}_0 \otimes \mathbf{I}) + \frac{1}{2}(\mathbf{C}_0\mathbf{W}^T\mathbf{W} \otimes \mathbf{I}).$$

To simulate the `RM` and `GiIrr` systems, correction terms (such as $\nabla \cdot \mathbf{B}(\theta)$) need to be computed. The correction terms are derived using the symbolic algebra toolbox in MATLAB. Since the perturbations are vectors of polynomials, the symbolic algebra toolbox can easily derive and efficiently evaluate the correction terms.

For the numerical experiments, we synthetically generate $m = 3$ signals, one of which is Laplace distributed, and two are distributed according to the squared hyperbolic secant distribution. The posterior distribution is $d = 9$ dimensional, there are a total of $N = 400$ data points, and the gradient is approximated by subsampling $n = 40$ data points per estimate. Since the posterior is nine-dimensional and highly multimodal, it is difficult to evaluate its marginal densities directly, i.e., without

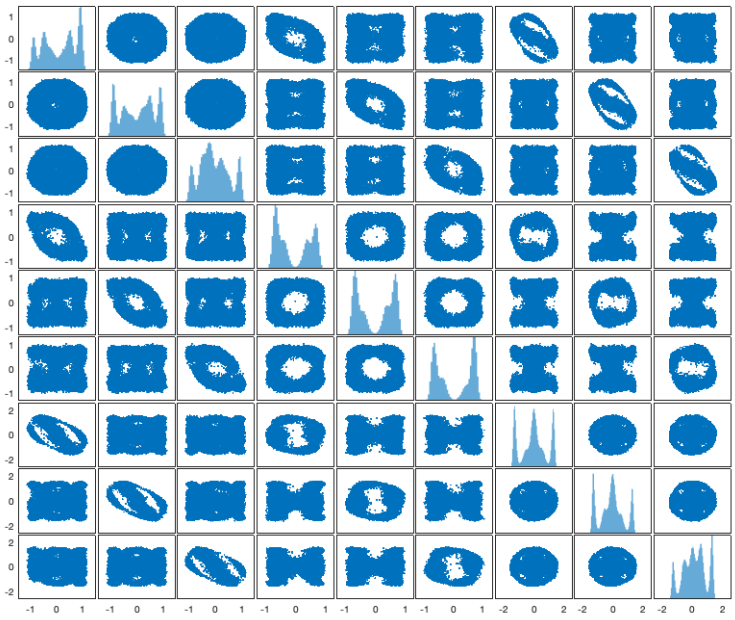


Figure 5-6: Posterior distribution sampled with standard Langevin with a deterministic gradient with $T = 10000$ and $h = 10^{-4}$. Notice that the system is very multimodal and non-Gaussian.

sampling. Instead, we establish a baseline reference density by simulating the standard Langevin dynamics with exact computation of the likelihood over all the data over $T = 10000$ with $h = 10^{-4}$. One- and two-dimensional marginals of this baseline posterior distribution are plotted in Figure 5-6. The two-dimensional marginals highlight the challenges of sampling from this posterior. In Figure 5-7, we plot trace plots of the $\mathbf{W}_{21}$ variable for each system. By visual inspection, we see that mixing is best for the geometry-informed irreversibly perturbed system. One can intuitively expect that with better mixing, the geometry-informed irreversibility should yield better estimation performance than the other systems. We assess this quantitatively below.

As in the previous example, we simulate the five systems and compute the asymptotic variances of two observables for each system. Each system is simulated independently 100 times up to time $T = 2000$ with $h = 2 \times 10^{-5}$. The smaller step size is to account for the additional stiffness irreversible perturbations introduce. Since the true

mean of the posterior distribution is unknown, and because standard sampling methods fail to adequately sample from the posterior distribution to get a reasonable estimate for the mean, we only plot the variance of the two observables with respect to K in Figure 5-8. To compute the asymptotic variance, we allot a burn-in time of $T_b = 20$. The observables we estimate are $\phi_1(\mathbf{W}) = \sum_{i,j} \mathbf{W}_{ij}$, and $\phi_2(\mathbf{W}) = \left(\sum_{i,j} \mathbf{W}_{ij}\right)^2$. The asymptotic variance numbers confirm that the faster mixing observed in the geometry-adapted irreversible perturbation does lead to a better sampling method. The values of the asymptotic variance are reported in Table 5.6. Notice that the geometry-informed irreversible perturbation far outperforms standard irreversibility applied to the reversible perturbation. When estimating the posterior mean, `GiIrr` yields an asymptotic variance that is more than 60 times smaller than that of `RM`.

	$\mathbb{E}[\text{AVar}_{\phi_1}]$	$\text{Std}[\text{AVar}_{\phi_1}]$	$\mathbb{E}[\text{AVar}_{\phi_2}]$	$\text{Std}[\text{AVar}_{\phi_2}]$
`LD`	81.76	23.88	50.21	17.92
`RM`	71.96	17.00	41.48	14.24
`Irr`	37.90	13.46	20.63	9.147
`RMIrr`	32.84	10.76	10.89	3.140
`GiIrr`	**1.182**	0.3881	**0.7419**	0.2391

Table 5.6: Asymptotic variance estimates for the ICA example.

5.4 Discussion

We presented a novel irreversible perturbation, `GiIrr`, that accelerates the convergence of Langevin dynamics. By introducing an irreversible perturbation that incorporates any given underlying reversible perturbation, which can also be interpreted as defining a Riemannian metric, we have shown through numerical examples that geometry-informed irreversible perturbations outperform those that are not informed as such. In the examples, we found that `GiIrr` seems to perform best when the target distribution is highly non-Gaussian.

Most of our numerical examples used stochastic gradients to cut down on compu-

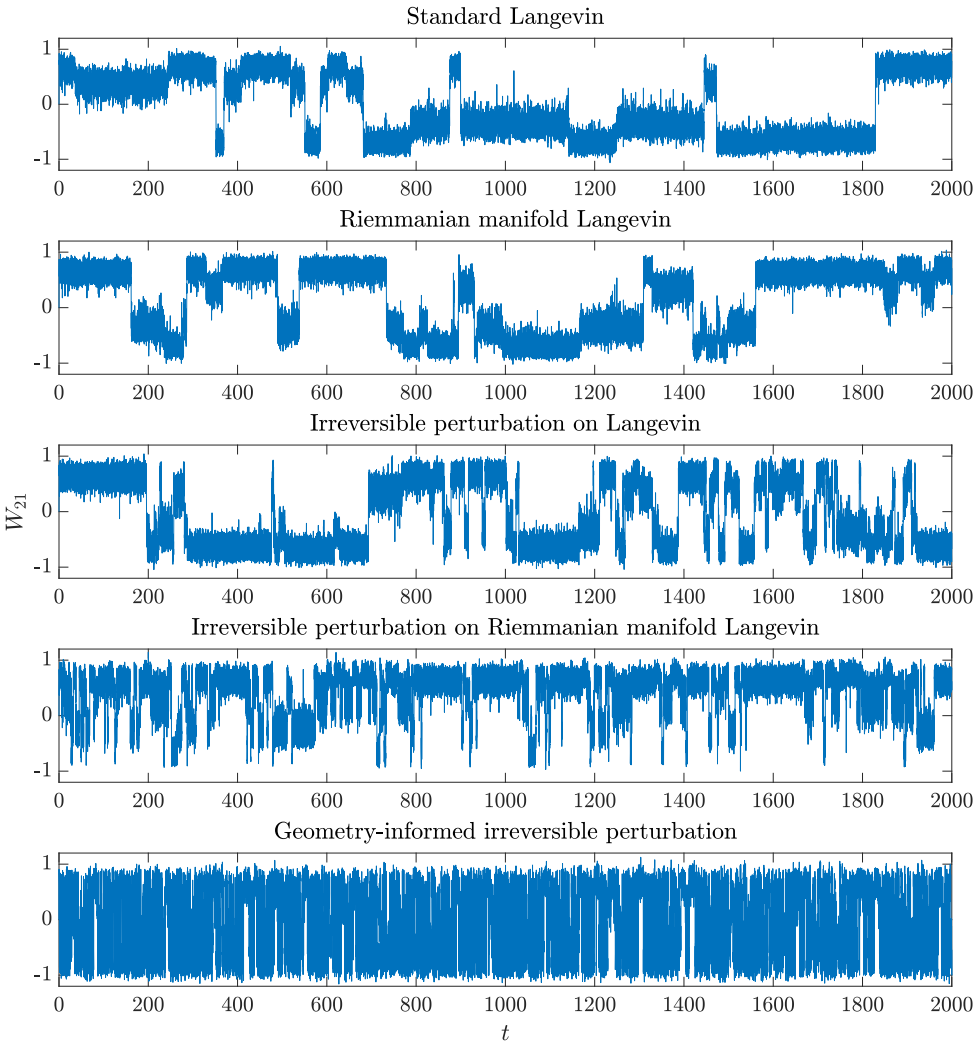


Figure 5-7: Trace plots of the W_{21} marginal.

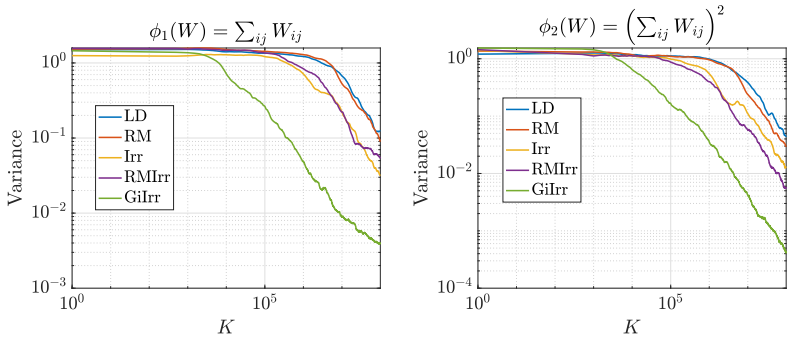


Figure 5-8: Variance of running average estimators

tational effort in sampling each trajectory. This demonstrates that SGLD can be used in conjunction with irreversibility for practical computations.

We also provided some analysis on how irreversibility interacts with discretization of the SDE systems. Irreversibility introduces additional stiffness into the system, which may lead to additional bias or variance in the estimator. For practical purposes, one can simply choose a small enough step size so that the asymptotic bias and variance are sufficiently small. At the same time, we note an example (see Appendix B) where the introduction of the irreversible term, once discretized, leads to no improvement in the long term average estimator.

Future work could study the use of novel integrators which circumvent stiffness. For example, [77] uses a multiscale integrator, but it is not readily adapted to the data-driven setting of Bayesian inference. Another direction for future work is to theoretically characterize the performance of the geometry-informed irreversible perturbation and to compare it with that of other perturbations. A starting point for such an analysis could be the general results of [97], in particular the large deviations Theorem 1 together with Propositions 2–4 therein. Preliminary investigation of this direction showed that it is a promising avenue for a theoretical investigation, but non-trivial work and a finer analysis are needed to demonstrate the effects of this class of irreversible perturbations. In the next chapter we turn our attention to reversible perturbations.

Chapter 6

Transport map unadjusted Langevin algorithm: guarantees and connections

6.1 Introduction

We continue developing methods for accelerating the convergence of Langevin dynamics. In this chapter we make two main contributions. We present the transport map unadjusted Langevin algorithm (TMULA), and show that when the transport map is chosen properly, standard results for fast convergence of ULA will apply even when the target distribution does not satisfy the usual assumptions of strong log-concavity. We also show that in the continuous setting, when the transport map is applied to standard Langevin dynamics, the result is an RMLD where the metric is defined by the Jacobian of the transport map. Connections to large deviations theory and variational formulations of Bayesian inference are made with the goal that these connections may spur future development of designing approximate maps that accelerate convergence of Langevin dynamics.

Transport maps are another recent approach to Bayesian computation [6, 80, 89, 109]. Transport maps provide a functional representation of complex random variables

by expressing them as the function of a simple random variable. We study the use of transport maps on the unadjusted Langevin algorithm (ULA). When the likelihood model is easy to compute (such as when the model is simple and the dataset is small), MCMC is used to obtain asymptotically unbiased samples from the target distribution. However with large datasets, one may not have the luxury of rejecting points. As a trade-off, ULA is used instead where one does not reject any samples, but contends with the resulting bias due to discretization. Recently, there has been much interest in the conditions a target distribution should satisfy to see rapid convergence of ULA [44, 126]. These conditions typically are that the log-target should have Lipschitz gradients and be strongly concave, or satisfy a log-Sobolev inequality. These conditions are not generally guaranteed, however we show that with the use of a transport map, we can obtain rapid convergence to the stationary distribution even when the distributions do not satisfy these properties. An invertible transport map can be applied to a target that does not satisfy typical conditions and create an ULA that still converges geometrically fast towards the target. Properties of the map only affect the constants in the convergence bounds and do not affect the rate of convergence. In particular, the *rate* of convergence is optimized when the pushforward of the target through the map is an isotropic Gaussian. At least in the context of ULA, this gives insight into why a transport maps should be constructed so that it pushes the target to a standard normal distribution.

Transport maps have also been found to improve MCMC algorithms [89] in which an approximate transport map from the target distribution to a simple Gaussian distribution was applied to Langevin dynamics of the pushforward of the target to create proposal distributions. As samples of the target were obtained, the map was updated. This can be viewed as a non-Gaussian extension of the adaptive Metropolis algorithm [55]. While the method was shown to be ergodic and empirically shown to produce better estimators for computing expectations with respect to posterior distributions, the precise reason for why the transport map accelerated convergence was not described. The authors of [89] conjectured that the map defined a Riemannian metric, however no proof was provided.

In this chapter, we also make connections between transport maps and the Riemannian manifold Langevin dynamics. In the previous chapter, we showed that it is beneficial to take the underlying reversible perturbation into account when applying an irreversible perturbations on an already reversibly perturbed Langevin dynamics. We previously assumed that the reversible perturbation was defined *a priori*. The design of reversible perturbations most often appears as the study of choosing the best Riemannian metric in the Riemannian manifold Langevin dynamics. In this chapter, we study a particular way to *parametrize* reversible perturbations using transport maps. Furthermore, we show that transport maps can also define geometry-informed Langevin algorithms. We also make connections to the theory of reversible perturbations and large deviations to explain why transport maps accelerates convergence. For future work, we state some conjectures relating our work to variational formulations of Bayesian inference.

6.1.1 Transport maps

The background we provide on transport maps are mainly based on [80, 89]. Transport maps provide functional representations of complex random variables in terms of simple random variables. Let X and Y be $\mathbb{R}^d$–valued random variables with probability measures μ_η and μ_π supported on all of $\mathbb{R}^d$, respectively. Let η and π be densities of μ_η and μ_π, respectively, with respect to the Lebesgue measure. We refer to η and π as the reference and target densities, respectively. A transport map $S : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ is a deterministic mapping such that the pushforward of μ_π under the map is equal to μ_η. That is, for any measurable set $A \subset \mathbb{R}^d$, $\mu_\pi(A) = \mu_\eta(S(A))$. The density π can be written in terms of the map S and the density η: $\pi(y) = \eta \circ S(y) \det \mathbf{J}_S(y)$, where $\mathbf{J}_S(y)$ is the Jacobian of S. While there are infinitely many mappings that can satisfy this relation, a well-studied restriction is the Knothe-Rosenblatt rearrangement where the map S is triangular, in which the i-th component of the map is dependent on the first i components of the argument, and monotone increasing, where the partial derivative of the i-th map is strictly positive [19, 80]. To be explicit, let $X = [x_1, \dots x_d]^\top$ and

$Y = [y_1, \ldots, y_d]^\top$, the map S has the form

$$S(Y) = \begin{bmatrix} S_1(y_1) \\ \vdots \\ S_d(y_1, y_2, \ldots, y_d) \end{bmatrix}, \tag{6.1}$$

where $\frac{\partial S_i}{\partial y_i}(Y) > 0$ for all i. There are several advantages for restricting to a map of this form. First, the pushforward density of π through S can be easily computed since the determinant of the Jacobian of S is the product of the diagonal entries. Second, the triangular structure of the map allows the inverse to be easily computed. Finding $S^{-1}(x)$ involves solving a sequence of d one-dimensional root finding problems, and since the i-th entry of the map is monotone with respect to x_i, the root is unique, and therefore the inverse map exists. In what follows, we denote the inverse map to be T.

If the density η were standard normal, then samples of the target π can be obtained by first procuring samples $x_i \sim \eta$ and computing the inverse map $y^{(i)} = S^{-1}(x^{(i)})$. How transport maps are constructed computationally is not a focus of this chapter, however, we provide some common approaches for approximating transport maps. One method for constructing maps from densities is to minimize the Kullback-Leibler (KL) divergence from π to the pushforward of η through the map T. We have the optimization problem

$$\min_{T \in \mathcal{T}} D_{KL}(T_\sharp \eta \| \pi) = \min_{T \in \mathcal{T}} D_{KL}\, \mathbb{E}_\eta \left[\log \left(\frac{\eta(x)}{\pi(T(y)) \det \mathbf{J}_T} \right) \right]. \tag{6.2}$$

Parametrizing the set of triangular, monotone increasing transport maps is explored in [6]. Another approach to constructing maps is through samples of π. This problem was considered in [89], in which the transport map is refined as more samples from the target were obtained.

In this chapter, we assume we have a map S built from a procedure such as this, except that $S_\sharp \pi$ may not be Gaussian. When $S_\sharp \pi$ is not Gaussian, using S directly for sampling will lead to bias, however when combined with a Langevin sampler, we can eliminate the bias. Moreover, we will show that when the Langevin sampler is

applied to a target density that has slow convergence, the use of the transport map can lead to faster rates of convergence.

6.2 Transport map-induced Riemannian metrics

In the previous chapter, we showed how certain perturbations to Langevin dynamics can accelerate its convergence and improve the performance of Langevin samplers. Transport maps have been shown to improve the performance of the Metropolis-adjusted Langevin algorithm, where the map is used in combination with Langevin dynamics to create non-Gaussian proposals for MCMC [89]. Here, we study how transport maps alter Langevin dynamics through the lens of reversible and irreversible perturbations. We show that the stochastic process induced by a transport map that acts on a standard Langevin dynamics is a Riemannian manifold Langevin system. Denote $\mathbf{J}_T$ and $\mathbf{J}_S$ to be the Jacobian matrices of maps T and S, respectively. Note that these matrices are lower triangular, and that by the inverse function theorem, note that $\mathbf{J}_T = \mathbf{J}_S^{-1}$.

The work of [89] present a method for accelerating the MALA by using transport maps to construct non-Gaussian proposals. Rather than using the transport map to pushforward the reference Gaussian distribution to the target, they instead map proposals of the reference (which are in the form of Equation (6.3)) and use those as proposals for sampling from π. They then noted that this proposal performed better than the standard MALA proposal and hypothesized that this is due to the fact that the transport map defines a Riemannian metric, and therefore induces a Riemannian manifold Langevin dynamics (RMLD). When the underlying metric is designed properly, the RMLD has been noted to have a faster rate of convergence to the stationary density than standard Langevin dynamics [53, 76, 134]. The authors conjectured that the Riemannian metric induced by the transport map is $\mathbf{J}_S^* \mathbf{J}_S$, but they did not definitively prove that this is the case [89]. RMLD has also been shown to be an example of a reversible perturbation on the original Langevin dynamics [97, 134]. The equivalent reversible perturbation is found by considering the matrix

$\mathbf{B} = (\mathbf{J}_S^* \mathbf{J}_S)^{-1}$. In the following proposition, we prove the conjecture formulated in [89]. Let $X(t)$ be defined as

$$\mathrm{d}X(t) = \nabla \log \eta(X(t))\mathrm{d}t + \sqrt{2}\mathrm{d}W(t) \tag{6.3}$$

Proposition 3. *(TM + LD = RMLD) Let X, Y be $\mathbb{R}^d$-valued random variables with densities $\eta(x)$ and $\pi(y)$, where η is a standard normal density. Let $T : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ be a twice-continuous, invertible map such that $\pi = T_\sharp \eta$. The diffusion process $T(X(t))$ where $X(t)$ evolves according to Equation (6.3) is equivalent to*

$$dY(t) = [\mathbf{B}(Y(t))\nabla \log \pi(Y(t)) + \nabla \cdot \mathbf{B}(Y(t))]dt + \sqrt{2\mathbf{B}(Y(t))}dW(t) \tag{6.4}$$

where $\mathbf{B} = (\mathbf{J}_S^* \mathbf{J}_S(Y))^{-1} = \mathbf{J}_T \mathbf{J}_T^*$, and $\sqrt{\mathbf{B}} = \mathbf{J}_S^{-1} = \mathbf{J}_T$.[1]

Proof. We first derive the SDE of the diffusion process $Z(t) = T(X(t))$. By Itô's lemma [85], the k–th component of $Z(t)$ is

$$\begin{aligned}
\mathrm{d}Z_k(t) =& \sum_{i=1}^{d} \frac{\partial T_k}{\partial x_i}\mathrm{d}X_i(t) + \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2 T_k}{\partial x_i \partial x_j}\mathrm{d}X_i(t)\mathrm{d}X_j(t) \\
=& \sum_{i=1}^{d} \frac{\partial T_k}{\partial x_i}\left(\frac{\partial}{\partial x_i}\log \eta(X(t))\mathrm{d}t + \sqrt{2}\mathrm{d}W_i(t)\right) \\
&+ \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2 T_k}{\partial x_i \partial x_j}\left(\frac{\partial}{\partial x_i}\log \eta(X(t))\mathrm{d}t + \sqrt{2}\mathrm{d}W_i(t)\right)\left(\frac{\partial}{\partial x_j}\log \eta(X(t))\mathrm{d}t + \sqrt{2}\mathrm{d}W_j(t)\right).
\end{aligned}$$

By the Itô calculus, note that $\mathrm{d}t \cdot \mathrm{d}t = \mathrm{d}t \cdot dW_i = 0$ and that $\mathrm{d}W_i \cdot \mathrm{d}W_j = \delta_{ij}\mathrm{d}t$. so we have that

$$\mathrm{d}Z_k(t) = \sum_{i=1}^{d} \frac{\partial T_k}{\partial x_i}\frac{\partial}{\partial x_i}\log \eta(X(t))\mathrm{d}t + \sum_{i=1}^{d} \frac{\partial^2 T_k}{\partial x_i^2}\mathrm{d}t + \sqrt{2}\sum_{i=1}^{d} \frac{\partial T_k}{\partial x_i}\mathrm{d}W_i(t).$$

[1] The divergence of a matrix is defined as $(\nabla \cdot \mathbf{B})_i = \sum_j \frac{\partial}{\partial x_j}\mathbf{B}_{ij}$.

Taken together, we may write

$$\mathrm{d}Z(t) = \mathbf{J}_T \nabla_X \log \eta(S(Z(t)))\mathrm{d}t + c(Z(t))\mathrm{d}t + \sqrt{2}\mathbf{J}_T \mathrm{d}W(t), \tag{6.5}$$

where

$$c_k(Z) = \sum_{i=1}^{d} \frac{\partial^2 T_k}{\partial x_i^2} = \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\partial^2 T_k}{\partial y_j \partial x_i} \cdot \frac{\partial T_j}{\partial x_i}. \tag{6.6}$$

The second equality is true by the multivariate chain rule and will be useful later.

As for the RMLD, we perturb $\{Y_t\}$ by the matrix $\mathbf{B} = \mathbf{J}_T \mathbf{J}_T^*$. Then after substituting $\pi(y) = \eta(S(y))|\mathbf{J}_S(y)|$, we have

$$\mathrm{d}Y(t) = \mathbf{J}_T \mathbf{J}_T^* \nabla_Y \log[\eta(S(Y(t)))\det(\mathbf{J}_S)]\mathrm{d}t + \nabla \cdot (\mathbf{J}_T \mathbf{J}_T^*)\mathrm{d}t + \sqrt{2}\mathbf{J}_T \mathrm{d}W(t). \tag{6.7}$$

First note that the diffusion term is equivalent to that of Equation (6.5), so we only need to compare the drift terms. Next, notice that

$$\frac{\partial}{\partial y_k} \log(\eta(S(Y))) = \sum_{i=k}^{d} \frac{\partial S_i}{\partial y_k} \frac{\partial}{\partial x_i} \log(\eta(S(Y)))$$

Therefore, $\nabla_Y \log \eta(S(Y)) = \mathbf{J}_S^* \nabla_X \log \eta(S(Y))$ and that $\mathbf{J}_T \mathbf{J}_T^* \nabla_Y \log \eta(S(Y)) = \mathbf{J}_T \nabla_X \log \eta(S(Y))$. This exactly matches the first part of the drift term in Equation (6.5).

For the divergence term, first note the following identity. Let $\mathbf{A}$ and $\mathbf{D}$ be $d \times d$ matrix-valued functions. Then observe that

$$\begin{aligned}
(\nabla \cdot \mathbf{AD})_k &= \sum_{j=1}^{d} \frac{\partial}{\partial y_j} (\mathbf{AD})_{kj} = \sum_{j=1}^{d}\sum_{i=1}^{d} \frac{\partial}{\partial y_j} \mathbf{A}_{ki}\mathbf{D}_{ij} \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\partial \mathbf{A}_{ki}}{\partial y_j}\mathbf{D}_{ij} + \mathbf{A}_{ki}\frac{\partial \mathbf{D}_{ij}}{\partial y_j} \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d} \left[\frac{\partial \mathbf{A}_{ki}}{\partial y_j}\mathbf{D}_{ij}\right] + (\mathbf{A}\nabla \cdot \mathbf{D})_k.
\end{aligned}$$

Applying this identity to $\mathbf{B} = \mathbf{J}_T\mathbf{J}_T^*$, we obtain

$$\begin{aligned}(\nabla \cdot \mathbf{J}_T\mathbf{J}_T^*)_k &= \sum_{i,j=1}^{d}\left[\frac{\partial \mathbf{J}_{ki}}{\partial y_j}\mathbf{J}_{ij}^*\right] + (\mathbf{J}_T\nabla \cdot \mathbf{J}_T^*)_k \\ &= \sum_{i,j=1}^{d}\left[\frac{\partial^2 T_k}{\partial y_j \partial x_i}\cdot\frac{\partial T_j}{\partial x_i}\right] + (\mathbf{J}_T\nabla \cdot \mathbf{J}_T^*)_k.\end{aligned}$$

Notice that the first term is exactly the Itô correction term! Hence, we only need to show that $\mathbf{J}_T\mathbf{J}_T^*\nabla \log\det \mathbf{J}_S + \mathbf{J}_T\nabla \cdot \mathbf{J}_T^* = 0$. To avoid issues with the chain rule relating Y and X, it suffices to show that $\nabla \log\det \mathbf{J}_S + \mathbf{J}_S^*\nabla \cdot (\mathbf{J}_S^*)^{-1} = 0$.

For the first term, observe that

$$\begin{aligned}\frac{\partial}{\partial y_k}\log\det \mathbf{J}_S &= \sum_{i,j=1}^{d}((\mathbf{J}_S^*)^{-1})_{ij}\left(\frac{\partial \mathbf{J}_S}{\partial y_k}\right)_{ij} \\ &= \sum_{i,j=1}^{d}\left(\frac{\partial^2 S_i}{\partial y_k \partial y_j}\right)((\mathbf{J}_S^*)^{-1})_{ij} \\ &= \sum_{i,j=1}^{d}\left(\frac{\partial \mathbf{J}_S}{\partial y_j}\right)_{ik}((\mathbf{J}_S^*)^{-1})_{ij}\end{aligned}$$

Letting $:j$ denote the j–th column vector, we see

$$\begin{aligned}(\mathbf{J}_S^*\nabla \cdot (\mathbf{J}_S^*)^{-1})_k &= -\left(\mathbf{J}_S^*\sum_{j=1}^{d}\left[(\mathbf{J}_S^*)^{-1}\frac{\partial \mathbf{J}_S^*}{\partial y_j}(\mathbf{J}_S^*)^{-1}\right]_{:j}\right)_k = -\left(\sum_{j=1}^{d}\left[\frac{\partial \mathbf{J}_S^*}{\partial y_j}(\mathbf{J}_S^*)^{-1}\right]_{:j}\right)_k \\ &= -\sum_{j=1}^{d}\left[\frac{\partial \mathbf{J}_S^*}{\partial y_j}(\mathbf{J}_S^*)^{-1}\right]_{kj} \\ &= -\sum_{j,i=1}^{d}\left(\frac{\partial \mathbf{J}_S^*}{\partial y_j}\right)_{ki}((\mathbf{J}_S^*)^{-1})_{ij}.\end{aligned}$$

which exactly the negative of the derivative of the log determinant of $\mathbf{J}_S$. We therefore conclude that Equation (6.5) exactly matches (6.7).

□

Note that the previous result only requires the map to be twice-continuously

differentiable and invertible. It is not dependent on the triangular structure of certain transport maps.

6.2.1 Transport maps induce geometry-informed irreversibility

Irreversible perturbations to Langevin dynamics are known to accelerate convergence to the equilibrium distribution by taking advantage of the anisotropy of the target distribution. A natural question to ask is, what is the output stochastic process of a transport map applied to a reference Langevin dynamics that has an irreversible perturbation? The following proposition states that the output is an geometry-informed irreversible perturbation applied to the Riemannian manifold Langevin dynamics derived in Proposition 1. The irreversible perturbation takes the Riemannian metric into account.

Proposition 4. *(`TM + Irr = GiIrr`) Define triangular, monotone transport map T be such that $T_\sharp \eta = \pi$. Let $X(t) \in \mathbb{R}^d$ evolve according to*

$$dX(t) = (\mathbf{I} + \mathbf{D})\nabla \log \eta(X(t))dt + \sqrt{2}dW(t)$$

where $\mathbf{D}$ is a skew-symmetric matrix. Then the stochastic process $Y(t) = T(X(t))$ evolves according to

$$\begin{aligned} dY(t) = [(\mathbf{B}(Y(t)) + \mathbf{C}(Y(t)))\nabla \log \pi(Y(t)) + \nabla \cdot (\mathbf{B}(Y(t)) + \mathbf{C}(Y(t)))]dt \\ + \sqrt{2\mathbf{B}(Y(t))}dW(t) \end{aligned} \tag{6.8}$$

where $\mathbf{B}(Y)$ is defined in Equation (6.4), and $\mathbf{C}(Y) = \mathbf{J}_T(Y)\mathbf{D}\mathbf{J}_T^*(Y) = (\mathbf{J}_S(Y)^*)^{-1}\mathbf{D}\mathbf{J}_S(Y)^{-1}$.

Proof. We apply Itô's formula and obtain

$$\mathrm{d}Y_k(t) = \sum_{i=1}^{d} \frac{\partial T_k}{\partial x_i}\left[\frac{\partial}{\partial x_i}\log\eta(X(t)) + (\mathbf{D}\nabla\log\eta(X(t)))_i\right]\mathrm{d}t + \sum_{i=1}^{d}\frac{\partial^2 T_k}{\partial x_i^2}\mathrm{d}t + \sqrt{2}\sum_{i=1}^{d}\frac{\partial T_k}{\partial x_i}\mathrm{d}W_i(t).$$

The first, third, and fourth terms are identical to the ones in Proposition 1 when irreversibility is not considered. We only need to address the second term. First note that

$$\begin{aligned}\mathbf{J}_T\mathbf{D}\nabla\log\eta(X(t)) &= \mathbf{D}\mathbf{J}_T^*\nabla_Y\log\eta(S(Y)) \\ &= \mathbf{J}_T\mathbf{D}\mathbf{J}_T^*\nabla_Y\log(\eta(S(Y))\det\mathbf{J}_S(Y)) - \mathbf{J}_T\mathbf{D}\mathbf{J}_T^*\nabla_Y\log\det\mathbf{J}_S(Y) \\ &= \mathbf{J}_T\mathbf{D}\mathbf{J}_T^*\nabla_Y\log\pi(Y) - \mathbf{J}_T\mathbf{D}\mathbf{J}_T^*\log\det\mathbf{J}_S(Y).\end{aligned}$$

We only need to show that the last term is equal to $\nabla\cdot\mathbf{J}_T\mathbf{D}\mathbf{J}_T^*$. From the proof of the previous proposition, first observe that

$$(\nabla\cdot\mathbf{J}_T\mathbf{D}\mathbf{J}_T^*)_k = \sum_{i,j=1}^{d}\left[\frac{\partial}{\partial y_j}(\mathbf{J}_T\mathbf{D})_{ki}\frac{\partial T_j}{\partial x_i}\right] + (\mathbf{J}_T\mathbf{D}\nabla\cdot\mathbf{J}_T^*)_k$$

and notice that $\mathbf{J}_T\mathbf{D}\nabla\cdot\mathbf{J}_T^* = -\mathbf{J}_T\mathbf{D}\mathbf{J}_T^*\nabla\log\det\mathbf{J}_S$. Therefore, we only need to show that the first term in the equal above is identically zero. We compute

$$\begin{aligned}\sum_{i,j=1}^{d}\left[\frac{\partial}{\partial y_j}(\mathbf{J}_T\mathbf{D})_{ki}\frac{\partial T_j}{\partial x_i}\right] &= \sum_{i,j,l=1}^{d}\left[\mathbf{D}_{li}\frac{\partial^2 T_k}{\partial y_j\partial x_l}\frac{\partial T_j}{\partial x_i}\right] \\ &= \sum_{i,l=1}^{d}\left[\mathbf{D}_{li}\frac{\partial^2 T_k}{\partial x_i\partial x_l}\right] \\ &= \sum_{i,l=1,i>l}^{d}\mathbf{D}_{li}\left[\frac{\partial^2 T_k}{\partial x_i\partial x_l} - \frac{\partial^2 T_k}{\partial x_l\partial x_i}\right] = 0.\end{aligned}$$

□

6.2.2 Transport maps preserve large deviations principles

Transport maps applied to Langevin dynamics lead to interesting implications and structures in the study of large deviations of empirical measures and estimators of ergodic averages of observables. We review some notions of large deviations theory for SDEs as presented in [97]. Let $\mathcal{P}(\mathbb{R}^d)$ denote the space of probability measures on $\mathbb{R}^d$.

Definition 2. *Let $\{\mu_t\}$ be a sequence of random probability measures. The sequence is said to satisfy a large deviations principle (LDP) with rate function $I : \mathcal{P}(\mathbb{R}^d) \to [0, \infty]$ if*

- *For all open sets $O \subset \mathcal{P}(\mathbb{R}^d)$,*

$$\liminf_{t\to\infty} \frac{1}{t} \log \mathbb{P}(\mu_t \in O) \geq - \inf_{\mu \in O} I(\mu);$$

- *For all closed sets $F \subset \mathcal{P}(\mathbb{R}^d)$,*

$$\limsup_{t\to\infty} \frac{1}{t} \log \mathbb{P}(\mu_t \in F) \leq - \inf_{\mu \in F} I(\mu);$$

- *The levels sets $\{\mu | I(\mu) \leq M\}$ are compact in $\mathcal{P}(\mathbb{R}^d)$ for all $M < \infty$.*

Intuitively, large deviations principles for the invariant measure quantify how fast an empirical distribution convergences to the invariant measure.

Theorem 4. *Consider an SDE with infinitesimal generator*

$$\mathcal{L} = \frac{1}{2} \nabla \cdot a(x) \nabla + b(x) \nabla$$

Suppose measures $\mu \in \mathcal{P}(\mathbb{R}^d)$ has density $p(x)dx = \mu(dx)$ and $p(x) \in \mathcal{C}^{2+\alpha}(\mathbb{R}^d)$ for $\alpha > 0$. Then the Donsker-Varadhan rate function is

$$I(\mu) = \frac{1}{8} \int_{\mathbb{R}^d} \frac{\nabla p(x)^* a(x) \nabla p(x)}{p(x)^2} d\mu(x) - \frac{1}{2} \int_{\mathbb{R}^d} \frac{b(x) \nabla p(x)}{p(x)} d\mu(x) + \frac{1}{2} \int_{\mathbb{R}^d} \nabla \phi(x) a(x) \nabla \phi(x) d\mu(x)$$

where $\phi(x)$ is a solution of the equation

$$\nabla \cdot [p(x)(b(x) + a(x)\nabla\phi(x))] = 0.$$

For standard Langevin dynamics, i.e., if $b(x) = \nabla \log \pi(x)$ and $a(x) = 2\mathbf{I}$, we have the rate function

$$I_o(\mu) = \frac{1}{4}\int_{\mathbb{R}^d} \left\| \nabla \log \frac{p(x)}{\pi(x)} \right\|^2 \mathrm{d}\mu(x). \tag{6.9}$$

For Riemannian manifold Langevin dynamics, the rate function is

$$I_{\mathbf{B}}(\mu) = \frac{1}{4}\int_{\mathbb{R}^d} \left(\nabla \log \frac{p(x)}{\pi(x)} \right)^* \mathbf{B}(x) \left(\nabla \log \frac{p(x)}{\pi(x)} \right) \mathrm{d}\mu(x). \tag{6.10}$$

Large deviations principles for estimators of ergodic averages of observables are related to LDP of empirical measures by the contraction principle. Consider some observable $f : \mathbb{R}^d \to \mathbb{R}$ and the estimator

$$f_t = \frac{1}{t}\int_0^t f(X_s)\mathrm{d}s.$$

The contraction principle allows us to concludes that f_t has an LDP with rate function

$$\tilde{I}_f(l) = \inf_{\mu \in \mathcal{P}(\mathbb{R}^d)} \left\{ I(\mu) : \int f(x)\mathrm{d}\mu(x) = l \right\}. \tag{6.11}$$

It is known that the asymptotic variance of the estimator f_t is related to the second derivative of the rate function as follows

$$\sigma^2(f) = \frac{1}{2\tilde{I}_f''(\bar{f})} \tag{6.12}$$

where $\bar{f} = \int f(x)\pi(x)\mathrm{d}x$. From these results, [97] conclude that if $\mathbf{B}(x) - \mathbf{I}$ is positive definite, then the estimator constructed using RMLD has lower variance than the one constructed with standard Langevin. Next we present a result that relates the rate function of the reference Langevin system with the RMLD induced by the transport

map.

Proposition 5. *Let η and π be continuously differentiable reference and target densities with infinite support on $\mathbb{R}^d$ and let T be a triangular, monotone transport maps such that $T_\sharp \eta = \pi$. Define map S such that $S_\sharp \pi = \eta$. Let $X(t)$ be the stochastic process that evolves according to SDE in Equation (6.3) and $Y(t)$ be the RMLD in Equation (6.4) and let $I^\eta(\mu)$ and $I^{\pi,\mathbf{B}}(\mu)$ be their rate functions, respectively. Then, $I^\eta(\mu) = I^{\pi,\mathbf{B}}(T_\sharp \mu)$.*

Proof. We compute:

$$\begin{aligned} I^{\pi,\mathbf{B}}(T_\sharp \mu) &= \frac{1}{4}\int \left(\nabla_Y \log \frac{p(S(y))\det \mathbf{J}_S}{\pi(y)}\right)^* \mathbf{J}_T \mathbf{J}_T^* \left(\nabla_Y \log \frac{p(S(y))\det \mathbf{J}_S}{\pi(y)}\right) \mathrm{d}T_\sharp \mu(x) \\ &= \frac{1}{4}\int \left(\mathbf{J}_T^* \nabla_Y \log \frac{p(S(T(x)))\det \mathbf{J}_S}{\pi(T(x))}\right)^* \left(\mathbf{J}_T^* \nabla_Y \log \frac{p(S(T(x)))\det \mathbf{J}_S}{\pi(T(x))}\right) T_\sharp \mathrm{d}\mu(x) \\ &= \frac{1}{4}\int \left\| \mathbf{J}_T^* \nabla_Y \log \frac{p(x)}{\pi(T(x))\det \mathbf{J}_T} \right\|^2 \mathrm{d}\mu(x). \end{aligned}$$

Notice that the denominator is exactly equal to η and from previous computations (of the multivariate chain rule of transport maps), and we have $\nabla_Y = \mathbf{J}_S^* \nabla_X$. Therefore we obtain,

$$I^{\pi,\mathbf{B}}(T_\sharp \mu) = \frac{1}{4}\int \left\| \nabla_X \log \frac{p(x)}{\eta(x)} \right\|^2 \mathrm{d}\mu(x) = I^\eta(\mu).$$

□

We can interpret this result as a precise description of the intuition that the Langevin algorithm involving the transport map induced RMLD somehow has the same convergence properties of the reference Langevin dynamics. If the reference is a standard normal and the map is constructed exactly, then the performance of the RMLD on the complex target density is the same as that of LD on the standard normal.

Remark 1. *The large deviations rate function for empirical measures relating to Langevin dynamics is closely related to the Fisher divergence [63]. The Fisher diver-*

gence from density p to density q defined on manifold E is

$$\mathcal{F}(q\|p) = \int_E \left\| \nabla \log \frac{q(x)}{p(x)} \right\|^2 q(x)\,dx. \tag{6.13}$$

Therefore, one can see that, for example,

$$I^\eta(\mu) = \frac{1}{4}\mathcal{F}(p\|\eta), \tag{6.14}$$

where p is the density of μ with respect to the Lebesgue measure.

This result can also be seen in terms of LDP for estimators of ergodic average of observables.

Corollary 2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an observable, then $\tilde{I}_f^{\pi,\mathbf{B}}(l) = \tilde{I}_{f\circ T}^{\eta}(l)$, and therefore the asymptotic variance of $(f \circ T)_t$ and f_t are equal for the reference LD and target RMLD, respectively.*

Proof. Observe that

$$\begin{aligned}
\tilde{I}_f^{\pi,\mathbf{B}}(l) &= \inf_{\mu}\left\{ I^{\pi,\mathbf{B}}(\mu) : \int f(x)\mathrm{d}\mu(x) = l \right\} \\
&= \inf_{\mu}\left\{ I^{\eta}(S_\sharp\mu) : \int f(x)\mathrm{d}\mu(x) = l \right\} \\
&= \inf_{\nu = S_\sharp\mu}\left\{ I^{\eta}(\nu) : \int f(x)\mathrm{d}T_\sharp\nu(x) = l \right\} \\
&= \inf_{\nu}\left\{ I^{\eta}(\nu) : \int (f\circ T)(x)\mathrm{d}\nu(x) = l \right\} \\
&= \tilde{I}_{f\circ T}^{\eta}(l).
\end{aligned}$$

□

6.3 Transport map unadjusted Langevin algorithm

In the previous section, we found that transport maps can be used to define Riemannian metrics and studied how they interact with large deviations-based approaches for

analyzing reversible perturbations. When designed properly, RMLD can be a powerful tool for accelerating the convergence of Langevin dynamics. A good choice of metric may not be of practical use if it is very expensive to evaluate. After discretization, each step of RMLD requires evaluations of 1) the inverse metric, 2) the square root of the metric, and 3) the divergence of the metric. Often the bottleneck to applying RMLD is the computation of the divergence, especially when the metric is difficult to compute already. The connection between transport maps and RMLD provides another approach to applying RMLD, in which we map a discretization of overdamped Langevin dynamics through the transport map. We call the resulting algorithm the *transport map unadjusted Langevin algorithm.*

We first show sufficient conditions on the target distribution π and map S which guarantees fast convergence of TMULA to stationarity. The main idea is that the transport map allows us to relax the strongly log-concave condition that is usually required for the target. Define η to be the pushforward of π through S, and let $U(x) = -\log S_\sharp \pi(x)$. We make the following assumptions:

Assumption 1. *The map S is ρ-strongly monotone. That is,*

$$\|S(z) - S(z')\| \geq \rho \|z - z'\|.$$

When the Jacobian is well-defined, this condition is equivalent to $\mathbf{J}_S \succeq \rho \mathbf{I}$.

Assumption 2. *The function $U(x)$ is m-strongly convex, with L-Lipschitz gradients. For all $x, y \in \mathbb{R}^d$, there exists m and L such that*

$$U(y) \geq U(x) + \langle \nabla U(x), y - x \rangle + \frac{m}{2}\|x - y\|^2, \quad \|\nabla U(x) - \nabla U(y)\| \leq L\|x - y\|.$$

The assumption that S is strongly monotone implies that the inverse map is Lipschitz. Since there always exists an S such that $S_\sharp \pi$ is standard normal, we know there is at least one such map that satisfies these two conditions.

We obtain samples from π by first applying ULA to η and then applying the map T on the trajectories of the Langevin system defined by η. We write the gradient of

$\log \eta$ in terms of π and S:

$$\begin{aligned}\nabla_X \log \eta &= \nabla_X \log \pi(T(X)) \det \mathbf{J}_T(X) \\ &= \mathbf{J}_T^* \nabla_Y \log \pi(T(X)) - \sum_{i=1}^{d} \mathbf{J}_T^* \nabla_Y \log \frac{\partial S_i}{\partial y_i}.\end{aligned} \tag{6.15}$$

We obtain TMULA, which is defined by the following stochastic process:

$$\begin{cases} X^{k+1} = X^k + h\mathbf{J}_S^*(Y^k)^{-1}\left[\nabla_Y \log \pi(Y^k) - \sum_{i=1}^{d}\left(\frac{\partial S_i}{\partial y_i}(Y^k)\right)^{-1} H_i(Y^k)\right] + \sqrt{2h}\xi^{k+1} \\ Y^{k+1} = T(X^{k+1}) \end{cases} \tag{6.16}$$

where $H_i(Y^k) = \left[\frac{\partial^2 S_i}{\partial y_1 \partial y_i}, \cdots, \frac{\partial^2 S_i}{\partial y_d \partial y_i}\right]^\top$, where $\xi^{k+1} \sim \mathcal{N}(0, \mathbf{I})$.

We now show that the TMULA converges in the 2-Wasserstein distance to the target distribution when these assumptions are met. The 2-Wasserstein distance between probability measure μ and ν is

$$\mathcal{W}_2^2(\mu, \nu) = \inf_{W : W_\sharp \mu = \nu} \int_{\mathbb{R}^d} \|x - W(x)\|^2 \mathrm{d}\mu. \tag{6.17}$$

Proposition 6. *Denote η^k and π^k to be the distributions of the discrete-time process X^k and Y^k at time step k, respectively. Let $h \in (0, \frac{1}{m+L})$ and $\kappa = \frac{2mL}{m+L}$. Then,*

$$\mathcal{W}_2^2(\pi^k, \pi) \leq \frac{2}{\rho^2}\left(1 - \frac{\kappa h}{2}\right)^k \left[\|x - x^\star\|^2 + \frac{d}{m}\right] + C_1(h) \tag{6.18}$$

where

$$C_1(h) = \frac{2L^2 d}{\kappa h}\left[h^2\left(\kappa^{-1} + h\right)\right]\left[2 + \frac{L^2 h}{m} + \frac{L^2 h^2}{6}\right]\left(1 - \left(1 - \frac{\kappa h}{2}\right)^k\right). \tag{6.19}$$

The proof is straightforward, we directly apply Theorem 5 from [44]. We relate the convergence of ULA on $S_\sharp \pi$ to that of π by relating the Wasserstein distance between η^k and η with the distance between π^k and π. Specifically we use the following lemma,

which is also proven in [59]. The ρ-strongly monotone condition on S is crucial to the proof.

Lemma 1.

$$\mathcal{W}_2^2(\pi^k, \pi) \leq \frac{1}{\rho^2}\mathcal{W}_2^2(\eta^k, \eta) \tag{6.20}$$

Proof. We use the fact that S is ρ-strongly monotone:

$$\begin{aligned}\mathcal{W}_2^2(\eta^k, \eta) = \mathcal{W}_2^2(S_\sharp \pi^k, S_\sharp \pi) &= \inf_W \int \|x - W(x)\|^2 \mathrm{d}S_\sharp \pi^k(x) \\ &= \inf_W \int \|S(x) - W \circ S(x)\|^2 \mathrm{d}\pi^k(x) \\ \geq \rho^2 \inf_W \int \|x - S^{-1} \circ W \circ S(x)\|^2 \mathrm{d}\pi^k(x).&\end{aligned}$$

Since W is such that $W_\sharp S_\sharp \pi^k = S_\sharp \pi$, this implies that $S_\sharp^{-1} W_\sharp S_\sharp \pi^k = \pi$, and we therefore have

$$\mathcal{W}_2^2(\pi^k, \pi) \leq \frac{1}{\rho^2}\mathcal{W}_2^2(\eta^k, \eta).$$

□

Based on this statement, the rate of convergence can be optimized by considering the term $r = 1 - \frac{\kappa h}{2}$ alone. Choosing h to be as large as possible, the rate is

$$r = 1 - \frac{mL}{(m+L)^2}. \tag{6.21}$$

We show this rate is optimized if and only if η were an isotropic Gaussian. Fix $m > 0$, we study the function $r(L) = 1 - \frac{mL}{(m+L)^2}$ for $L \in [m, \infty)$. Taking the derivative with respect to L yields

$$\frac{\partial r}{\partial L} = \frac{m(L-m)}{(m+L)^3}, \tag{6.22}$$

which is always positive for $L > m$. Thus, the greatest rate of convergence is attained

when $L = m$, which implies that $\nabla^2 U(x) = \mathbf{I}$, and therefore implies that η is identically a standard normal. At least in the context of TMULA, this result supports the intuition that constructing a map S such that η is as close to an isotropic normal as possible is best.

6.3.1 Simple numerical example

To briefly demonstrate the differences between TMULA and a direct discretization of RMLD with the transport map defining the reversible perturbations, we provide the following simple numerical example. We consider a very thin banana distribution defined as follows

$$\log \pi(y) = -\frac{1}{2 \cdot 100^2} y_1^2 - \frac{1}{2}(y_2 + 2y_1^2 - 200)^2. \tag{6.23}$$

The transport map that takes samples from this distribution to a standard normal is

$$S(y) = \begin{bmatrix} y_1/10 \\ y_2 + 2y_1^2 - 200 \end{bmatrix}. \tag{6.24}$$

We simulate the standard Langevin dynamics with Euler-Maruyama, the transport map ULA, and a direct discretization of RMLD with the reversible perturbation defined by the transport map for a simulation length of $T_b = 100$ with time step $h = 5 \times 10^{-4}$. The results are produced in Figure 6-1. We compare the results to samples directly produced from the transport map. These figures show that TMULA and RMLD accelerates convergence to the stationary distribution because they are reversible perturbations applied to the Langevin process.

Visually note, however, that RMLD seems to have difficulties staying on the very narrow valley defined by the log-density. This is likely due to the stiffness of the directly discretized RMLD system. In contrast, TMULA is the better able to stay in the valley since the reference process (an Ornstein-Uhlenbeck process) is not stiff.

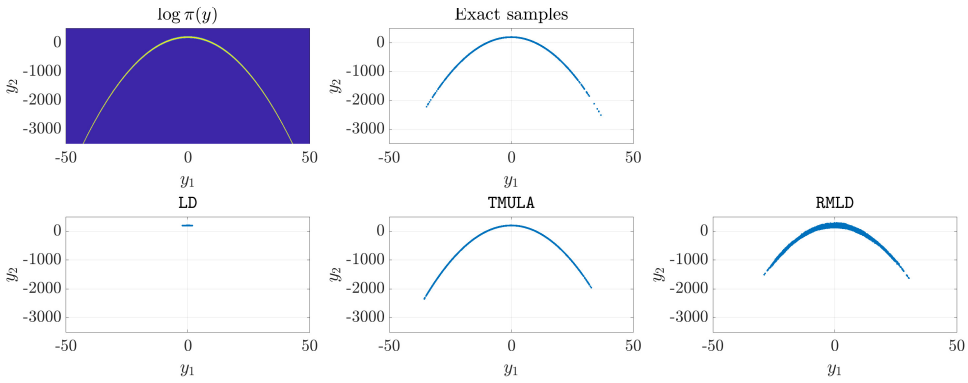


Figure 6-1: Banana distribution. Top left shows the log density, top right shows exact samples produced by a transport map. Bottom left shows single trajectory of Langevin dynamics, bottom middle shows single trajectory of TMULA, bottom right shows single trajectory of direct discretization of RMLD. Simulation length $T_s = 100$, time step $h = 5 \times 10^{-4}$.

6.4 Discussion and future work

In this chapter we studied the use of transport maps for accelerating the convergence of Langevin dynamics. For the unadjusted Langevin algorithm, it is clear that transport maps expands the space of distributions for which there are guarantees of rapid convergence. We have also made interesting connections to the Riemannian manifold Langevin dynamics as well, however, these results do not easily translate into computational notions that we can employ for practical use yet. Indeed, while an exact map will accelerate convergence of the invariant distribution, having access to the exact map also means that one does not have to resort to Langevin dynamics at all. In particular, we hope this work can potentially lead to interesting new objective functions for constructing approximate transport maps that optimally accelerates convergence of Langevin dynamics.

We showed the connections between transport maps and Riemannian manifold Langevin dynamics. Not only do transport maps provide a way to parametrize a certain class of Riemannian metrics, they also provide another way to *discretize* RMLD. There are, however, many more interesting directions we can investigate with this connection at hand. We discuss two possible future directions in this section.

6.4.1 Connections to variational formulations of Bayesian inference

We restrict ourselves to target distributions that arise from Bayesian statistics. In the context of Bayesian inference, another approach to evaluating the efficacy of the RMLD is to study variational formulations of the Bayesian update [117]. One can cast the Bayesian posterior distribution as the solution to a variational problem on the space of distributions with the Wasserstein distance. Following the presentation of [117], we let $\pi_0(y)$ be the prior density and let $l(y) \propto \exp(-\phi(y))$ be the likelihood. Let J_{KL} be a functional defined on the space of probability densities on $\mathbb{R}^d$

$$J_{KL}(\rho) := \mathrm{D}_{\mathrm{KL}}(\rho \| \pi_0) + \int_{\mathbb{R}^d} \phi(y)\rho(y)\mathrm{d}y. \tag{6.25}$$

The minimizer of this functional over the space of probability distribution is equal to the posterior distribution $\pi(y) \propto l(y)\pi_0(y)$.

The following result relates the convexity of the optimization problem to the Riemannian metric. For each $x \in \mathbb{R}^d$, define the inner product $g_x(u, v) = \langle \mathbf{G}(x)u, v \rangle$, where $u, v \in \mathbb{R}^d$. Using a framework for defining gradient flows in the space of probability distributions with the 2-Wasserstein distance as the metric [2], the authors of [117] show that the gradient flow derived from (6.25) is the overdamped Langevin dynamics. The convexity of this optimization problem, which defines the rate of convergence of the flow to the minimizer in the KL-divergence, with different metrics defined by the Riemannian metric is derived in [117].

Proposition 7 (Proposition 3 and Theorem 10 from [117])**.** *The geodesic convexity of the variational formulation for Bayesian inference with metric* $\mathbf{G}$ *is*

$$\lambda_{\mathbf{G}} = \inf_{v,y \in \mathbb{R}^d} \mathrm{Ric}_g(v, v) - \mathrm{Hess}_g \log[\pi_g(y)](v, v) \tag{6.26}$$

where $g(v, v) = 1$*,* $\pi_g(y)$ *is the density of the posterior with respect to the measure defined on the manifold,* $\mathrm{Ric}_g(v, v)$ *is the Ricci curvature tensor and* $\mathrm{Hess}_g(v, v)$ *is the Hessian tensor defined with respect to the metric.*

For our particular choice of metric, we compute that $\log \pi_g(y) = -\frac{1}{2}\|S(y)\|^2$. We conjecture that when the exact transport map is used, $\lambda_{\mathbf{G}} = 1$ identically. To prove this, we need to compute the Ricci curvature and Hessian terms above. The Christoffel symbols of the Riemannian manifold with metric $\mathbf{G}$ is defined

$$\Gamma_{ij}^l = \frac{1}{2}\left(\frac{\partial}{\partial x_j}\mathbf{G}_{ki} + \frac{\partial}{\partial x_i}\mathbf{G}_{kj} - \frac{\partial}{\partial x_k}\mathbf{G}_{ij}\right)\mathbf{G}_{lk}^{-1}. \tag{6.27}$$

And the Ricci curvature tensor and Hessian tensors are

$$(\mathrm{Ric}_g)_{ij} = \frac{\partial \Gamma_{ij}^l}{\partial x_l} - \frac{\partial \Gamma_{il}^l}{\partial x_j} + \Gamma_{ij}^l\Gamma_{lk}^k - \Gamma_{il}^k\Gamma_{jk}^l. \tag{6.28}$$

$$(\mathrm{Hess}_g \log[\pi_g(y)])_{ij} = \frac{\partial^2}{\partial y_i \partial y_j}\left(\frac{1}{2}\|S(y)\|^2\right) - \Gamma_{ij}^l\frac{\partial}{\partial y_l}\frac{1}{2}\|S(y)\|^2. \tag{6.29}$$

Computing these terms is generally quite difficult, however it would be useful to see how the transport map implies a metric and changes the properties of the resulting optimization problem.

We consider a slightly different formulation of the above proposition that does not use the definition of the Ricci curvature tensor.

Proposition 8 (Proposition 3 in [117])**.** *Let $F \in \mathcal{C}^2(\mathbb{R}^d)$ and $\mu(du) \propto \exp(-F(u))du$. The sharp constant λ for which D_{KL} is λ-geodesically convex in the g-Wasserstein distance is equal to*

$$\lambda_{\mathbf{G}} := \inf_{x\in\mathbb{R}^d} \Lambda_{min}\big(\mathbf{G}^{-1/2}(\mathbf{B} + \nabla^2 F - \mathbf{C})\mathbf{G}^{-1/2}\big)$$

where $\nabla^2 F$ is the usual Euclidean Hessian matrix of F, $\mathbf{B}$ is the matrix with coordinates

$$\mathbf{B}_{ij} = \frac{\partial \Gamma_{ij}^l}{\partial x_l} - \Gamma_{il}^k\Gamma_{jk}^l$$

and C is the matrix with coordinates

$$\mathbf{C}_{ij} = \Gamma_{ij}^l\frac{\partial F}{\partial x_l}.$$

We conjecture that with any transport map T such that $T_{\#}\eta = \pi$, where π is the target distribution and η is the pullback of π through T, the constant λ of RMLD with target π is the same as that of LD on pullback η.

Conjecture 1. *Let λ_η be such that*

$$\lambda_\eta := \inf_{x\in\mathbb{R}^d} \Lambda_{min}\big(-\nabla^2 \log \eta\big)$$

and λ_π be equal to that of $\lambda_{\mathbf{G}}$ above with $\mathbf{G} = \mathbf{J}_S^*\mathbf{J}_S$ and $F = -\log\pi$. Then $\lambda_\eta = \lambda_\pi$.

First observe that

$$\nabla_Y^2 \log \pi(y) = \nabla_Y^2 \log \eta(S(y)) + \nabla_Y^2 \log\det \mathbf{J}_S(y).$$

The first term is equal to the following

$$\nabla_Y^2 \log \eta(S(y)) = \mathbf{J}_S^*(y)\nabla_X^2 \log\eta(x)\mathbf{J}_S(y) + \sum_{k=1}^{n} \frac{\partial}{\partial y_k} \log\eta(S(y))\nabla_Y^2 S_k(y).$$

After pre- and post- multiplying the first term by $\mathbf{G}^{-1/2}$, we have exactly the term inside of λ_η. Therefore, we only need to show that

$$\frac{\partial \Gamma_{ij}^l}{\partial y_l} - \Gamma_{il}^k\Gamma_{jk}^l + \frac{\partial}{\partial y_k}\log\eta(S(y))\frac{\partial^2 S_k}{\partial y_i \partial y_j} + \Gamma_{ij}^l \frac{\partial}{\partial x_l}\log\eta(S(y)) + \Gamma_{ij}^l\frac{\partial}{\partial y_l}\log\det\mathbf{J}_S = 0.$$

If η were standard normal, this result would that the exact transport map that maps the standard normal to the target distribution will define a Riemannian metric such that the convexity of the optimization problem on the target is the same as that of the normal distribution.

Proving this result may be useful for other types of Langevin dynamics-based Bayesian inference methodology beyond RMLD. For example, sampling methodologies based on interacting Langevin diffusions use an empirically estimated covariance matrix to define the reversible perturbation [51]. Like for RMLD, Transport maps may be a generalization of the covariance matrix used in the interacting Langevin diffusions methodology.

6.4.2 Creating new objective functions for designing Riemannian metrics

While we have shown that Riemannian metrics can be parametrized through transport maps, we did not describe a way to use the theory of reversible perturbations to inform how to design transport maps that will result in metrics that accelerate convergence of Langevin dynamics. The large deviations approach for analyzing reversible perturbations provides a rate function which is used to measure the quality of a Langevin sampler, but using rate function to define an objective function for constructing transport maps does not seem straightforward. While an exact transport map is desirable, having it implies one no longer needs Langevin dynamics to sample from the target distribution. For future research, it is desirable to find some way of finding good Riemannian metrics based on approximate transport maps.

Chapter 7

Conclusion and future work

The goal of this thesis was to exhibit the symbiotic nature of recent advances in computational statistics and computational dynamics. We conclude by providing a unified discussion of all the major ideas presented here. We highlight common themes and challenges the chapters share with each other. From there, we specify future research directions brought up from ideas in this thesis.

7.1 Sampling *for* stochastic dynamical systems

In the first part, we demonstrated how state-of-the-art tools in computational dynamical systems can be used to devise efficient statistical tools for characterizing rare events in stochastic dynamical systems and quantifying their probabilities. Advances in numerical methods for Koopman operators provide new ways of computing the eigenfunctions of the operator of the Kolmogorov backward equation, thereby enabling cheap, approximate solutions to the KBE. By expressing the indicator function over rare regions of interest in terms of these eigenfunctions, we are able to approximate zero-variance importance sampling estimators and optimal implementations of multilevel splitting estimators. These eigenfunctions are learned from sampling typical trajectories of the dynamical system at different initial conditions.

Our techniques are generally agnostic to the exact Koopman numerical method applied; however, the main computational bottleneck for our approach is how well

Koopman eigenfunctions can be learned from trajectory data and how cheaply the eigenfunctions and their gradients can be evaluated throughout the domain. While atypical or rare trajectories that exhibit the rare event are not needed to successfully devise an efficient estimator, having training points near typical rare event paths improve estimator performance. In the examples on which we demonstrated our approaches, we found that learning a few eigenfunctions can reduce the variance of the rare event estimator by four orders of magnitude.

The ultimate goal of this work is to devise a fully *black-box* approach for estimating rare event probabilities and characterizing their mechanisms, i.e., a rare event sampling algorithm that solely depends on trajectory data. To describe a path to achieve this goal, we highlight limitations of our current work and present possible strategies to circumvent those limitations in the pursuit of this totally black-box algorithm.

- **Current limitations of Koopman numerical methods.** The importance sampling and multilevel splitting methodologies presented in this thesis critically depends on how well the stochastic Koopman eigenfunctions can be learned from data. We applied extended dynamic mode decomposition and its variants, in which a basis is specified *a priori* and the action of the stochastic Koopman operator on the basis is projected onto the basis to obtain a finite dimensional representation of the operator. While this approach enables the fast evaluation of the approximate eigenfunctions and their gradients, the algorithm often produces spurious eigenvalues and eigenfunctions, which need to be identified and removed before applying our rare event sampling algorithm. Furthermore, when the basis is not chosen properly, the quality of the eigenfunctions can often be quite poor. To alleviate this problem, one extension of this approach is to adaptively learn the basis simultaneously or iteratively while learning the eigenfunctions. Neural nets have been employed mostly due to their expressiveness [74, 113]. However all of these approaches suffer from the curse of dimensionality since the number of basis functions needed to effectively express the eigenfunctions will grow exponentially with dimension.

To circumvent the issues that arise from basis-dependent methods, we may also consider basis-free methods, such as Hankel DMD [3, 31, 129] or extensions of the diffusion maps algorithm [8, 9]. These methods approximate the value of the eigenfunctions at the training points, and at least in the case of Hankel DMD, can be scaled to high dimensions since the computational cost scales with the size of the amount of trajectory data. These methods also have convergence guarantees. For Hankel DMD, it was shown that, assuming an invariant subspace could be identified, the approximated eigenfunctions will converge to the true eigenfunctions in the limit of infinite data over a long trajectory [129]. In the case of diffusion maps-based methods, there are guarantees where the stochastic matrix produced by the diffusion maps algorithm will converge to the generator of the diffusion process for noisy gradient systems in the limit of infinite data.

Both methods still implicitly depend on a basis of initial observables. Given an initial set of observables, Hankel DMD produces a dictionary of observables through the repeated application of the system's stochastic Koopman operator. The idea behind this approach is based on Krylov subspace methodology, in which the basis produced by the powers of the operator applied to an initial basis function is able to quickly approximate an invariant subspace. One drawback of this approach is that one still has to define a good initial set of basis functions for the method to work in practical settings.

On the other hand, the diffusion maps algorithm constructs a data-driven basis using trajectory data. Given data, the diffusion maps algorithm yields an orthonormal basis composed of eigenfunctions of the gradient system defined by the log of the data-producing density [9, 10]. Moreover, it has been shown that this particular basis is the optimal orthonormal basis as it minimizes the Dirichlet energy which allows the Koopman eigenfunctions to be learned robustly. Specifically, the estimators for learning the expansion coefficients will have minimal variance [8]. A drawback of this approach is that the basis is constructed from a kernel, which means that it may be scalable in high

dimensions (say for more than dimension 10). A future direction of research may be to see how diffusion maps can be combined with Hankel DMD methods. For example, a small initial set of basis functions can be learned from the diffusion maps algorithm and Hankel DMD is applied to generate more eigenfunctions.

While the basis-dependent approach to Koopman numerical methods may often produce spurious eigenfunctions, they produce eigenfunctions whose gradients can be quickly evaluated throughout the state space. This is advantageous for our rare event sampling algorithm as both the importance sampling and splitting methods require repeated evaluation of the eigenfunctions or their gradients. Applying importance sampling with eigenfunctions learned from Hankel DMD or the diffusion maps approach may be impractical as the gradients will need to be estimated from eigenfunction evaluations from training points, which would have very large variance. A more viable approach is to only consider multilevel splitting and investigate the use of high dimensional interpolation in evaluating the eigenfunctions. Notwithstanding the drawbacks of Koopman numerical methods, further investigation of these methods will be crucial in the development of a black-box rare event sampling algorithm.

- **Adaptive exploration of the state space.** Ultimately, we are pursuing a rare event sampling algorithm with only the ability to simulate from the black-box of some stochastic dynamical process. In theory, our method only depends on the sample trajectories. What parts of the state space the sample trajectories explore is important in developing rare event sampling algorithms. While trajectories that exhibit the rare event do not need to be included in the dataset, the training data do need to at least cover the various transition pathways for our algorithm to be effective. In high dimensions, it is difficult to have training data cover all parts of the state space, therefore it is best to have a strategy to adaptively explore the state space. For example, we may draw inspiration from [27], in which a diffusion maps algorithm is applied to adaptively explore a free energy landscape to find saddle points and transitions

between local minima. This approach was only studied in the context of gradient systems, and was not used in the estimation of probabilistic quantities. It may be interesting to see how our work can be combined with these ideas.

- **Projecting the indicator function on to the approximated stochastic Koopman eigenfunctions.**

  Once the eigenfunctions are estimated, another challenge is to project the indicator function over the rare event onto the eigenfunctions. In Chapter 2, we saw that to apply importance sampling properly, the approximate function needs to be strictly positive. We addressed this issue in a relatively arbitrary way, in which we simply added a positive constant to the approximation so that it was strictly positive over the training points. For future work, it may be interesting to investigate more rigorous ways for positive regression.

  This problem may also be of interest for multilevel splitting. While multilevel splitting does not need the approximation to be positive, we have not yet thoroughly investigated how our approach works for splitting when the approximation is not corrected.

- **Characterizing the types of rare events for which our approach will be effective.** Throughout the Koopman rare event sampling methodology, there is an implicit assumption that indicator functions over rare regions of interest can be expressed in terms of the Koopman eigenfunctions. Of course, this is not true in general, so it would useful if we could characterize the types of rare event problems our approach would be effective.

7.2 Sampling *by* stochastic dynamical systems

In the second part of the thesis, we showed how dynamical systems can be used to sample from complex probability distributions. While Langevin samplers such as the unadjusted Langevin algorithm (ULA) exhibits exponential convergence in the Wasserstein distance, the class of distributions that satisfy the conditions required

for fast convergence is rather restrictive. We presented three approaches for sampling with stochastic dynamical systems.

Langevin samplers rely on ergodicity, i.e., the property that the empirical average of a function over a single trajectory will eventually equal the expectation the function with respect to the target distribution. However, the empirical average may converge very slowly. One way to resolve this problem is to create a dynamical system that can sample from target distributions in finite time. In Chapter 4, we re-interpreted our approach to importance sampling for diffusion processes to exactly sample from a general class of probability distributions with Gaussian tails. We presented an approach to building controlled SDEs whose *finite*-time marginal distribution will match the target distribution. By wisely choosing a linear SDE as the reference process, we are able to devise exact solutions of the Kolmogorov backward equation in terms of eigenfunctions of the Ornstein-Uhlenbeck operator. By expressing the ratio of the target density and the marginal density of the uncontrolled system at some finite time T, we are able to derive the optimal control that produces a controlled process that samples from the target distribution by only solving a static optimization problem. This was in contrast to approaches based on the full Schrödinger bridge problem [7], where usually some sort of infinite-dimensional optimization problem needs to be solved. While the approach is elegant in its presentation, its practical implementation to high-dimensional problems in Bayesian inference requires certain challenges to be addressed.

- **Expressing the ratio of the target and reference densities in terms of the OU eigenfunctions.** Like the rare event sampling problem, one of the primary challenges of this method is to efficiently expand functions in terms of the KBE operator's eigenfunctions. In our formulation, we can pursue a least-squares regression approach, however the issue is that the approximating function needs to be strictly positive so that the control can be properly applied. The approach we take results in a highly non-convex objective function whose evaluation needs to be estimated with sample points. It would also be interesting to see how the properties of this optimization problem will change with the

alternate formulations we presented in the conclusion of Chapter 4.

- **Informed ways for choosing the reference process and versions of the algorithm.** Even when the likelihood ratio can be sparsely expressed in terms of the eigenfunctions, we found that the quality of the approximation is very sensitive the choice of the eigenfunctions (which is dependent on the choice of reference process). In our examples, the reference processes were chosen in an *ad hoc* fashion. To make the algorithm more practical and robust, it would be beneficial to devise a way to automatically choose and refine the choice of the reference process.

- **Beyond linear systems: other choices of the reference process.** We chose linear stochastic systems to be the reference process because their Fokker-Planck (FPE) and Kolmogorov backward equations (KBE) can be solved analytically through eigenfunction expansions. It would be interesting to see if there are nonlinear SDEs whose FPE and KBE could be solved exactly. This would expand the class of approximating distributions and potentially make our approach richer and more robust.

- **Solving the Schrödinger bridge problem.** The controlled SDEs formulation we presented in this thesis is the solution to a special case of the Schrödinger bridge problem (SBP) when the initial condition is deterministic. In the conclusion of Chapter 4, we proposed a few different special cases of the SBP that may merit further investigation. For future work, it would be interesting to see if our formulations can be further developed to solve the full Schrödinger bridge problem. Finding more efficient ways of solving the SBP will be impactful for problems in data assimilation and optimal control.

The other approach we took to improve the performance of Langevin samplers is the design of novel reversible and irreversible perturbations. In contrast to controlled SDEs, the reference process is the overdamped Langevin dynamics and the dynamics are changed so that the system converges more quickly to the stationary distribution,

but still in infinite time. The most well-known instantiation of reversible perturbations is the Riemannian manifold Langevin dynamics, which is known to accelerate convergence when the difference between Riemannian metric and the identity is positive definite. Meanwhile, irreversible perturbations are simple and cheap to find and compute, and in continuous-time, guarantees accelerated convergence. In Chapter 5 we proposed and demonstrated a geometry-informed irreversible perturbation that accelerated the convergence of Riemannian manifold Langevin dynamics more than if standard irreversibility were applied alone. This geometry-informed irreversibility took the underlying Riemannian metric into account when designing the irreversible perturbation.

- **Analysis on when geometry-informed irreversibility performs better.** While our numerical examples showed the promise of state-dependent irreversibility through numerical experiments, further detailed investigation needs to be done to understand under what conditions does geometry-informed irreversibility accelerate convergence of Langevin dynamics and how to design it. While there is previous work analyzing irreversible perturbations with large deviations theory, applying this mode of analysis for geometry-informed irreversibility has not been fully explored.

- **Synthesis and analysis of new discretization schemes for irreversible perturbations.** While irreversibility is known to always improve the performance of Langevin samplers in continuous-time, this property does not hold in discrete-time. In Appendix B we show how the variance of the estimator can increase when irreversibility is introduced. Proposing discretizations other than Euler-Maruyama can potentially alleviate some of these difficulties.

- **Computational studies for trading off sophisticated perturbations with producing more samples.** While perturbations (both reversible and irreversible) accelerates convergence of Langevin dynamics, they incur additional computational cost when applied. An open question is how does this additional computational cost trade off with the improvements in sampling performance? In

Chapter 5, we saw that for the Bayesian logistic regression example the reversible perturbations were quite expensive to compute, and the improvements in performance were relatively marginal when compared with unperturbed Langevin dynamics. It would be beneficial if one could determine if applying perturbations is worthwhile compared to simulating longer trajectories in standard Langevin *a priori*.

In Chapter 6, we studied reversible perturbations in the context of transport maps. Given an invertible map and a target distribution, we define a reference Langevin process on the pullback of the distribution through the map. When the inverse the map is applied to the reference Langevin process, the output is a Riemannian manifold Langevin process with a reversible perturbation that is defined by the Jacobian of the map. When the map is applied to an irreversibly perturbed reference process, the result is a geometry-informed irreversibly perturbed Langevin system on the target distribution. The transport maps not only parametrize reversible perturbations, but they also provide new way to discretize Riemannian manifold Langevin dynamics. With this connection, we introduced the transport map unadjusted Langevin algorithm, which is easier to simulate than a discretized Riemannian manifold Langevin dynamics when the map is available. This connection, while interesting, requires further development on how to construct the transport map.

- **Devising alternative objective functions for finding transport maps and reversible perturbations.** Our analysis shows that an exact transport map that takes a standard Gaussian to the target distribution optimizes the rate of convergence of the TMULA in the 2-Wasserstein distance. Having the optimal transport map renders the use of Langevin dynamics moot since we can use the map directly to produce samples from the target. One goal is to produce new objective functions for finding approximate transport maps that can be directly linked to the improved sampler performance.

- **Further properties of transport maps and reversible perturbations.**

Information geometry appears in contexts beyond Riemannian manifold Langevin dynamics. The metric defined by transport maps are another way of encoding information geometry. Further research could involve using this new perspective for other kinds of Bayesian methodology, such as the construction of interacting particle systems [51]. In Chapter 6, we highlighted the variational formulation of Bayesian inference, in which different kinds of Riemannian metrics can increase the convexity of the variational problems associated with Bayesian inference. Connections such as these could result in even more interesting algorithms in Bayesian computation.

In this chapter, we showed that transport maps can defined Riemannian metrics. Another interesting avenue of exploration is studying the necessary conditions a Riemannian metric should satisfy so that it corresponds with a transport map. This would enable TMULA for new classes of Riemannian metrics.

Appendix A

Computing eigenfunctions of the multidimensional Ornstein–Uhlenbeck operator

We discuss approaches to computing eigenfunctions of the Ornstein–Uhlenbeck (OU) operator in more than two dimensions. While the spectrum of the OU operator and theoretical properties of its eigenfunctions have been well characterized in previous research, the practical computation of general eigenfunctions has not been resolved. We review special cases for which the eigenfunctions can be expressed exactly in terms of commonly used orthogonal polynomials. Then we present a tractable approach for computing the eigenfunctions in general cases and comment on its dimension dependence.

A.1 Introduction

The Ornstein–Uhlenbeck (OU) operator naturally arises in many fields. In stochastic differential equations (SDEs), the OU operator is the generator of the Ornstein–Uhlenbeck semigroup, which describes the evolution of statistics OU *processes*, which are linear time-homogeneous SDEs [90]. Eigenfunctions of the OU operator also appear in Koopman operator analysis of linear stochastic dynamical systems, as the stochastic

Koopman operator for linear SDEs has the same eigenfunctions as the OU operator [31]. These eigenfunctions have been useful in perturbation analysis of Fokker–Planck equations for nonlinear SDEs [70]. Recently, the eigenfunctions have been shown to be useful in constructing importance sampling schemes for rare event simulation [137]. The OU process is also used to model dynamical phenomena in financial mathematics [84, 125] and neuroscience [47, 99].

Properties of the spectrum and eigenfunctions of the OU operator have been thoroughly explored in the literature. For example, the spectrum has been computed exactly, and many theoretical properties of the eigenfunctions—such as the fact that they are polynomials and are complete in certain weighted L^p spaces—have been established [82]. There are, however, applications in which one needs to directly work with the eigenfunctions [70, 137]. The exact form of the eigenfunctions has only been recorded in limited special cases, and a comprehensive approach to computing the eigenfunctions, in general, has not been found by the authors. In this note, we describe certain cases in which the multidimensional OU eigenfunctions can be represented compactly in terms of commonly used orthogonal polynomials. Then we outline a direct way of computing them in a more general setting. This note is targeted towards those who are looking for methods to *exactly* compute the eigenfunctions of the OU operator for *general* diagonalizable drift and diffusion matrices, in *arbitrary dimensions*.

A.2 Theory and special cases

A.2.1 Notation and problem setting

Let $\mathbf{A}$ and $\mathbf{B}$ be $d \times d$ and $d \times r$ real-valued matrices, respectively, with $d \geq r$, and define $\mathbf{Q} = \frac{1}{2}\mathbf{B}\mathbf{B}^\top$, where $^\top$ denotes the matrix transpose. Below, $\overline{\lambda}$ will denote the complex conjugate, * will denote the conjugate transpose, and $\langle u, v \rangle = u^* v$ will be the inner product. Assume that the eigenvalues of $\mathbf{A}$ have strictly negative real parts, and that none of the left eigenvectors of $\mathbf{A}$ are contained in the kernel of $\mathbf{B}^\top$. We also

assume that $\mathbf{A}$ is diagonalizable; $\mathbf{B}$ may be rank-deficient.[1] We study the computation of the eigenfunctions on $L^p(\nu)$ for $p > 1$, where ν is the invariant probability measure associated with the linear system of the operator. The existence of a nondegenerate invariant measure ν is guaranteed by the assumptions on $\mathbf{A}$ and $\mathbf{B}$ [82]. The OU operator $\mathcal{A}$ is given by

$$\mathcal{A}\psi = \langle \mathbf{A}x, \nabla\psi \rangle + \operatorname{Tr} \mathbf{Q}\nabla^2\psi = \sum_{i=1}^{d} (\mathbf{A}x)_i \frac{\partial \psi}{\partial x_i} + \sum_{i,j=1}^{d} \mathbf{Q}_{ij} \frac{\partial^2 \psi}{\partial x_i \partial x_j}. \tag{A.1}$$

In the context of stochastic differential equations, the OU operator is the infinitesimal generator of the OU process, which is a time-homogeneous linear SDE,

$$\mathrm{d}X_t = \mathbf{A}X_t \mathrm{d}t + \mathbf{B}\mathrm{d}W_t\,, \tag{A.2}$$

where W_t is a standard d-dimensional Brownian motion.

The spectrum of the Ornstein–Uhlenbeck operator and its associated semigroup has been well studied (for example, see [82, 11, 78]). Previous research has characterized the eigenfunctions of the self-adjoint OU operator, which corresponds to the case when $\mathbf{A}$ is self-adjoint and shares the same eigenvectors as $\mathbf{B}$. In this case, the eigenfunctions are the tensorized Hermite polynomials [90]. In $d = 2$ dimensions, if $\mathbf{A}$ has only complex eigenvalues and is normal (i.e., $\mathbf{A}^\top\mathbf{A} = \mathbf{A}\mathbf{A}^\top$), the eigenfunctions are the so-called Hermite-Laguerre-Itô (HLI) polynomials [24]. In general the OU operator is not self-adjoint, so we cannot appeal to the spectral theory of self-adjoint operators to prove the existence of eigenvalues. Nevertheless, the seminal work of [82] shows that, under mild conditions, the OU operator has a pure point spectrum in $L^p(\nu)$ for $1 < p < \infty$, where ν is the stationary measure of the OU process. Moreover, [82] shows that the eigenfunctions form a complete basis in $L^p(\nu)$ for $1 < p < \infty$, the eigenfunctions are all polynomials, and that the eigenvalues and eigenfunctions are the same for all $1 < p < \infty$. We summarize these facts by recalling the following propositions from [82].

[1] When $\mathbf{B}$ is rank-deficient, this leads to the case where the Ornstein–Uhlenbeck operator is hypoelliptic [82].

Proposition 9 ([82, Theorem 3.1]). *Let $-\lambda_1, \ldots, -\lambda_l$ be the distinct eigenvalues of $\boldsymbol{A}$, where $\lambda_k > 0$ for all k. Then the spectrum of $\mathcal{A}$ is given by*

$$\left\{ -\sum_{k=1}^{l} n_k \lambda_k : n_k \in \mathbb{N} \right\}.$$

Moreover, the linear span of the eigenfunctions of $\mathcal{A}$ is dense in $L^p(\nu)$.

Proposition 10 ([82, Proposition 3.1]). *Suppose that u is in the domain of $\mathcal{A}$ and satisfies $(\gamma - \mathcal{A})u = 0$ for some $\gamma \in \mathbb{C}$. Then u is a polynomial of degree less than or equal to $|Re(\gamma)/s(A)|$, where $s(A) = \sup_k\{Re(\lambda_k)\}$. That is, the eigenfunctions of the OU operator are polynomials.*

In [71], the authors describe the generalized form of the OU eigenfunctions in terms of ladder operators. Given a seed eigenfunction, repeated application of the ladder operators generates other eigenfunctions. While compact in its mathematical formulation, the approach is not easily amenable to practical computations. To make computing eigenfunctions tractable, we represent the OU operator as a matrix acting in some chosen basis of polynomials. Since it is known that the eigenfunctions of the OU operator are polynomials, an exact matrix representation of the OU operator on some finite dimensional vector space of polynomials is possible [82].

While the pure point spectrum of the OU operator on $L^p(\nu)$ spaces with $p > 1$ is known explicitly, there is no explicit expression for the eigenfunctions in general. In [82], the authors showed that for $p > 1$ the spectrum of the OU operator is the same as that of

$$\mathcal{L}\psi \coloneqq \langle x, \mathbf{A}^\top \nabla \psi \rangle = \sum_{k=1}^{d} x_k \big(\mathbf{A}^\top \nabla \psi\big)_k, \tag{A.3}$$

regardless of the form of the diffusion term. In Section A.4, we will show how the eigenfunctions of $\mathcal{L}$ in fact comprise a judicious choice of basis for computing the eigenfunctions in general. The following lemma will be useful later when converting the OU eigenvalue problem into a matrix eigenvalue problem.

Lemma 1. *Let $\mathbf{A} \in \mathbb{R}^{d\times d}$ be diagonalizable and full rank. Let f_i be a left eigenvector of $\mathbf{A}$ with eigenvalue $-\lambda_k$, i.e., $f_k^*\mathbf{A} = -\lambda_k f_k^*$. Let $\mathbf{n} \in \mathbb{N}_0^d$ be a d-dimensional multi-index of nonnegative integers. The eigenfunctions of the operator $\mathcal{L}\psi = \langle x, \mathbf{A}^\top \nabla\psi\rangle$ are*

$$\psi_{\mathbf{n}}(x) := \prod_{k=1}^{d} \psi_{n_k}(x) = \prod_{k=1}^{d} \langle x, f_k\rangle^{n_k} \tag{A.4}$$

with eigenvalues

$$\mu_{\mathbf{n}} = -\sum_{k=1}^{d} n_k\lambda_k. \tag{A.5}$$

Proof. Observe that

$$\begin{aligned}
\left\langle x, \mathbf{A}^\top\nabla\prod_{k=1}^{d}\psi_{n_k}(x)\right\rangle &= \left\langle x, \mathbf{A}^\top\sum_{j=1}^{d} n_j\langle x, f_j\rangle^{n_j-1} f_j \prod_{k\neq j}^{d}\langle x, f_k\rangle^{n_k}\right\rangle \\
&= \sum_{j=1}^{d} n_j\langle x, \mathbf{A}^\top f_j\rangle\langle x, f_j\rangle^{n_j-1}\prod_{k\neq j}^{d}\langle x, f_k\rangle^{n_k} \\
&= \sum_{j=1}^{d} -n_j\lambda_j\prod_{k=1}^{d}\langle x, f_k\rangle^{n_k} \\
&= -\left(\sum_{j=1}^{d} n_j\lambda_j\right)\psi_{\mathbf{n}}(x).
\end{aligned}$$

□

A.3 Hermite and Hermite-Laguerre-Itô polynomials

In this section we review the definitions of the Hermite and HLI polynomials, and some of their relevant properties.

A.3.1 Hermite polynomials

There are many ways to define the probabilists' Hermite polynomials. The most relevant characterization for this note is the Hermite differential equation, which is an eigenvalue problem of the form

$$-x\phi_n'(x) + \phi_n''(x) = \mu_n \phi_n(x). \tag{A.6}$$

The solutions to this differential equation are the Hermite polynomials $\phi_n(x) = \mathrm{He}_n(x)$ with eigenvalues $\mu_n = -n$ for $n \in \mathbb{N}_0$.

The Hermite polynomials (like any other univariate orthogonal polynomials) satisfy a three-term recurrence relation:

$$\mathrm{He}_{n+1}(x) = x\mathrm{He}_n(x) - n\mathrm{He}_{n-1}(x). \tag{A.7}$$

Furthermore, derivatives of the Hermite polynomials can be expressed in terms of other, lower-order, Hermite polynomials as

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathrm{He}_n(x) = n\mathrm{He}_{n-1}(x).$$

A.3.2 Hermite-Laguerre-Itô polynomials

The Hermite-Laguerre-Itô (HLI) polynomials are bivariate orthogonal polynomials first studied by Itô in his study of multiple complex-valued Itô integrals. The definition of the HLI polynomials used in this note is from [24]. A more comprehensive collection of the properties of these polynomials can also be found there. For integers m, n and $(x, y) \in \mathbb{R}^2$, the polynomials are

$$J_{m,n}(z, \overline{z}) = \begin{cases} (-1)^n n!\, z^{m-n} L_n^{m-n}(z\overline{z}, \rho), & m \geq n \\ (-1)^m m!\, \overline{z}^{n-m} L_m^{n-m}(z\overline{z}, \rho), & m < n \end{cases}$$

where $z = x + iy$ and $L_k^\alpha(x, \rho)$ are the generalized Laguerre polynomials defined by the Rodrigues formula

$$L_n^\alpha(x, \rho) = \frac{\rho^n}{n!} x^{-\alpha} e^{\frac{x}{\rho}} \frac{\mathrm{d}^n}{\mathrm{d}x^n} \left(e^{-\frac{x}{\rho}} x^{n+\alpha} \right), \quad n \in \mathbb{N}.$$

The first six Hermite-Laguerre-Itô polynomials for $\rho = 1$ are

$$J_{0,0} = 1,\ J_{1,0} = x + iy,\ J_{0,1} = x - iy$$
$$J_{1,1} = -(x^2 + y^2) + 1,\ J_{2,0} = (x + iy)^2,\ J_{0,2} = (x - iy)^2.$$

Like the Hermite polynomials, the derivatives of HLI polynomials can be written in terms of other HLI polynomials. Defining $z = x + iy$, we have,

$$\frac{\partial}{\partial z} J_{m,n}(z, \bar{z}) = m J_{m-1,n}(z, \bar{z})$$
$$\frac{\partial}{\partial \bar{z}} J_{m,n}(z, \bar{z}) = n J_{m-1,n}(z, \bar{z}).$$

The following crucial result from [24] shows that $J_{m,n}$ are the OU eigenfunctions

Proposition 11 ([24], Theorem 2.6)**.** *The Hermite-Laguerre-Itô polynomials satisfy*

$$\left[\overline{\lambda} z \frac{\partial}{\partial z} + \lambda \overline{z} \frac{\partial}{\partial \overline{z}} + 2\sigma^2 \frac{\partial}{\partial z \partial \overline{z}} \right] J_{m,n}(z, \bar{z}; \rho) = \mu_{m,n} J_{m,n}(z, \bar{z}; \rho)$$

where $\lambda = -a + ib$, $\rho = \sigma^2/a$, *and* $\mu_{m,n} = -(m + n)a + i(m - n)b$.

A.3.3 Special cases

The eigenfunctions of $\mathcal{A}$ are well-known for certain special cases. We outline some of these cases here.

A and B are self-adjoint and simultaneously diagonalizable

Here we study the case where **A** and **B** are self-adjoint and simultaneously diagonalizable. Then the eigenvalue problem is decomposable into d one-dimensional eigenvalue

problems, each of which is a Hermite differential equation. The relationship between the Hermite polynomials and the OU operator with $\mathbf{A} = \mathbf{B} = \mathbf{I}$ has been well-studied (see, e.g., [11, 78, 90] and the references therein). The extension to the present scenario is straightforward. The eigenvalues of $\mathbf{A}$ are real and the eigenvectors are orthogonal. Suppose we have $\mathbf{A}e_k = -\lambda_k e_k$ and $\mathbf{B}e_k = \sigma_k e_k$ for $\lambda_k, \sigma_k > 0$, with $\langle e_j, e_k \rangle = \delta_{jk}$. We first show that univariate Hermite polynomials defined in the direction of each of the eigenvectors are eigenfunctions. That is, we make the *ansatz* that

$$\phi_k(x) = g(\langle x, e_k \rangle) \tag{A.8}$$

and show that g can be expressed in terms of a Hermite polynomial. The gradient and Hessian of this function are

$$\nabla \phi_k(x) = g'(\langle x, e_k \rangle) e_k, \quad \nabla^2 \phi_k(x) = g''(\langle x, e_k \rangle) e_k e_k^\top, \tag{A.9}$$

so the OU operator applied to $\phi_k(x)$ yields

$$\mathcal{A}\phi_k(x) = \langle x, \mathbf{A}^\top e_k \rangle g'(\langle x, e_k \rangle) + \frac{1}{2} \operatorname{Tr}\big[\mathbf{B}\mathbf{B}^\top e_k e_k^\top\big] g''(\langle x, e_k \rangle).$$

This yields the eigenvalue problem,

$$-\lambda_k \langle x, e_k \rangle g'(\langle x, e_k \rangle) + \frac{1}{2}\sigma_k^2 g''(\langle x, e_k \rangle) = \mu_k g(\langle x, e_k \rangle). \tag{A.10}$$

Recall that the probabilist's Hermite polynomials $\mathrm{He}_n(x)$ solve the Hermite differential equation $-x\mathrm{He}_n'(x) + \mathrm{He}_n''(x) = -n\mathrm{He}_n(x)$. Therefore, notice that if $g(\langle x, e_k \rangle) = \mathrm{He}_{n_k}\left(\sqrt{\frac{2\lambda_k}{\sigma_k^2}}\langle x, e_k \rangle\right)$ for some $n_k \in \mathbb{N}_0$, we then have

$$\begin{aligned} -\lambda_k \sqrt{\frac{2\lambda_k}{\sigma_k^2}} \langle x, e_k \rangle \mathrm{He}_{n_k}'\left(\sqrt{\frac{2\lambda_k}{\sigma_k^2}}\langle x, e_k \rangle\right) + \lambda_k \mathrm{He}_{n_k}''\left(\sqrt{\frac{2\lambda_k}{\sigma_k^2}}\langle x, e_k \rangle\right) \\ = -n_k \lambda_k \mathrm{He}_{n_k}\left(\sqrt{\frac{2\lambda_k}{\sigma_k^2}}\langle x, e_k \rangle\right). \end{aligned} \tag{A.11}$$

In the next section, we will see that products of different univariate eigenfunctions defined in the directions of the eigenvectors of $\mathbf{A}$ are eigenfunctions of $\mathcal{A}$. Specifically,

$$\phi_{\mathbf{n}}(x) = \prod_{k=1}^{d} \phi_{n_k}(x) = \prod_{k=1}^{d} \mathrm{He}_{n_k}\left(\sqrt{\frac{2\lambda_k}{\sigma_k^2}}\langle x, e_k\rangle\right), \tag{A.12}$$

is an eigenfunction with eigenvalue $\mu_{\mathbf{n}} = -\sum_{k=1}^{d} n_k\lambda_k$.

A is normal and A, B are simultaneously diagonalizable

Now we consider the case where $\mathbf{A}$ is normal, i.e., $\mathbf{A}\mathbf{A}^\top = \mathbf{A}^\top\mathbf{A}$, but not necessarily self-adjoint. For this case, it is possible for $\mathbf{A}$ to have complex eigenvalues. In [24], for an OU operator with

$$\mathbf{A} = \begin{bmatrix} -a & b \\ -b & -a \end{bmatrix}, \text{ and } \mathbf{B} = \sigma\mathbf{I},$$

the OU eigenfunctions are found to be the Hermite-Laguerre-Itô (HLI) polynomials. The HLI polynomials are

$$J_{m,n}(z, \bar{z}; \rho) = \begin{cases} (-1)^n n!\, z^{m-n} L_n^{m-n}(z\bar{z}, \rho), & m \geq n \\ (-1)^m m!\, \bar{z}^{n-m} L_m^{n-m}(z\bar{z}, \rho), & m < n \end{cases}$$

where $L_k^\alpha(z, \rho)$ are the generalized Laguerre polynomials, $\rho = \sigma^2/a$, and $z = x_1 + ix_2$. The OU eigenvalues in this case are $\mu_{m,n} = -(m+n)a + i(m-n)b$. [24] also generalizes this result to d dimensions, for even d, when the matrix $\mathbf{A}$ is normal but only has complex eigenvalues. Similar to the self-adjoint case, the eigenfunctions are simply products of the HLI polynomials on each of the eigenspaces.

We now explicitly write the eigenfunctions for general normal matrices $\mathbf{A}$ and for self-adjoint matrices $\mathbf{B}$ that share the same eigenspace as $\mathbf{A}$. The latter conditions imply that $\mathbf{B}$ is only has real eigenvalues. While the expression follows simply from previous results, to our knowledge no previous work has explicitly computed these eigenfunctions.

When $\mathbf{A}$ has both real and complex eigenvalues, the eigenfunctions are products of Hermite and HLI polynomials. Suppose $\mathbf{A}$ has l eigenspaces, with l' real eigenspaces and $l-l'$ complex eigenspaces; that is, $\mathbf{A}$ has l' real eigenvalues and $l-l'$ pairs of complex eigenvalues. Let f_i denote a unit left eigenvector of $\mathbf{A}$ with eigenvalue $-\lambda_k$; that is, $f_k^*\mathbf{A} = -\lambda_k f_k^*$. Let the first l' eigenvalues be real and the next $l-l'$ eigenvalues come in complex conjugate pairs. To be clear, for complex eigenvalues, we write $\lambda_k = a_k - ib_k$. Let $\mathbf{B}$ be such that $\mathbf{B}f_k = \sigma_k f_k$, where $\sigma_k > 0$. Note that $l'+2(l-l') = d$. Let $\mathbf{n} \in \mathbb{N}_0^d$ be a multi-index defined as $\mathbf{n} = (n_1, \ldots, n_{l'}, n_{(l'+1)1}, n_{(l'+1)1}, \ldots, n_{l1}, n_{l2})$. Then the eigenfunction of the corresponding OU operator is

$$\phi_{\mathbf{n}} = \prod_{k=1}^{l} \phi_{n_k}(x) = \prod_{k=1}^{l'} \mathrm{He}_{n_k}\left(\sqrt{\frac{2\lambda_k}{\sigma_k^2}}\langle x, f_k\rangle\right) \cdot \prod_{k=l'+1}^{l} J_{n_{k1},n_{k2}}\left(\sqrt{2}\langle x, f_k\rangle, \sqrt{2}\,\overline{\langle x, f_k\rangle}; \rho_k\right) \tag{A.13}$$

with eigenvalue $\mu_{\mathbf{n}} = \sum_{k=1}^{l'} -n_k\lambda_k - \sum_{k=l'+1}^{l}[(n_{k1}+n_{k2})a_k - i(n_{k1}-n_{k2})b_k]$, and $\rho_k = \sigma_k^2/a_k$. We show that Equation (A.13) is indeed an OU eigenfunction. We first compute the following two expressions:

$$\begin{aligned}
\langle x, \mathbf{A}^\top \nabla\phi_{\mathbf{n}}(x)\rangle &= \sum_{k=1}^{l} \langle x, \mathbf{A}^\top \nabla\phi_{n_k}(x)\rangle \prod_{j=1,k\neq j}^{l} \phi_{n_j}(x) \\
\frac{1}{2}\operatorname{Tr}\mathbf{B}\mathbf{B}^\top\nabla^2\phi_{\mathbf{n}}(x) &= \frac{1}{2}\sum_{k=1}^{l} \operatorname{Tr}\mathbf{B}\mathbf{B}^\top\nabla^2\phi_{n_k}(x) \prod_{j=1,k\neq j}^{l} \phi_{n_j}(x) \\
&\quad + \sum_{k>j}^{l} \operatorname{Tr}\mathbf{B}\mathbf{B}^\top\nabla\phi_{n_k}(x)\nabla\phi_{n_j}(x)^\top \prod_{k'=1,k'\neq k\neq j}^{l} \phi_{n_{k'}}(x).
\end{aligned}$$

The gradient of ϕ_{n_k} is

$$\nabla\phi_{n_k}(x) = \begin{cases} \sqrt{\dfrac{2\lambda_k}{\sigma_k^2}}\mathrm{He}'_{n_k}\left(\sqrt{\dfrac{2\lambda_k}{\sigma_k^2}}\langle x, f_k\rangle\right) f_k \text{ if } 1 \le k \le l' \\ \sqrt{2}\dfrac{\partial J_{n_{k1},n_{k2}}}{\partial z_k} f_k + \sqrt{2}\dfrac{\partial J_{n_{k1},n_{k2}}}{\partial \overline{z}_k}\overline{f}_k \text{ if } l'+1 \le k \le l, \end{cases}$$

where $z_k = \sqrt{2}\langle x, f_k\rangle$. The Hessian of ϕ_{n_k} is

$$\nabla^2\phi_{n_k} = \begin{cases} \frac{2\lambda_k}{\sigma_k^2}\mathrm{He}''_{n_k}\left(\sqrt{\frac{2\lambda_k}{\sigma_k^2}}\right) f_k f_k^\top \text{ if } 1 \le k \le l' \\ 2\frac{\partial^2 J_{n_{k1},n_{k2}}}{\partial z_k^2} f_k f_k^\top + 2\frac{\partial^2 J_{n_{k1},n_{k2}}}{\partial \overline{z}_k^2}\overline{f}_k\overline{f}_k^\top + 4\frac{\partial^2 J_{n_{k1},n_{k2}}}{\partial z_k \partial \overline{z}_k} f_k\overline{f}_k^\top \text{ if } l'+1 \le k \le l. \end{cases}$$

By the normality of $\mathbf{A}$, the left eigenvectors of are orthonormal, so $\mathrm{Tr}[\mathbf{B}\mathbf{B}^\top f_k f_k^*] = \langle \mathbf{B}^\top f_k, \mathbf{B}^\top f_k\rangle = \sigma_k^2$, and $\mathrm{Tr}[\mathbf{B}\mathbf{B}^\top f_k f_j^*] = \langle \mathbf{B}^\top f_j, \mathbf{B}^\top f_k\rangle = \sigma_k\sigma_j\langle f_j, f_k\rangle = 0$. For cases where f_k is complex, i.e., when $l'+1 \le k \le l$, we also have $\mathrm{Tr}[\mathbf{B}\mathbf{B}^\top f_k f_k^\top] = \langle \mathbf{B}^\top\overline{f}_k, \mathbf{B}^\top f_k\rangle = \sigma_k^2\langle \overline{f}_k, f_k\rangle = 0$.

Next observe that for $1 \le k \le l'$, we have

$$\begin{aligned} \langle x, \mathbf{A}^\top\nabla\phi_{n_k}(x)\rangle + \frac{1}{2}\mathrm{Tr}\,\mathbf{B}\mathbf{B}^\top\nabla^2\phi_{n_k}(x) =& -\lambda_k\sqrt{\frac{2\lambda_k}{\sigma_k^2}}\langle x, f_k\rangle \mathrm{He}'_{n_k}\left(\sqrt{\frac{2\lambda_k}{\sigma_k^2}}\langle x, e_k\rangle\right) \\ &+ \lambda_k\mathrm{He}''_{n_k}\left(\sqrt{\frac{2\lambda_k}{\sigma_k^2}}\langle x, e_k\rangle\right) \\ =& -n_k\lambda_k\mathrm{He}_{n_k}\left(\sqrt{\frac{2\lambda_k}{\sigma_k^2}}\langle x, e_k\rangle\right). \end{aligned}$$

For $l'+1 \le k \le l$, we appeal to Proposition 11 in the Appendix to obtain

$$\begin{aligned} \langle x, \mathbf{A}^\top\nabla\phi_{n_k}(x)\rangle + \frac{1}{2}\mathrm{Tr}\,\mathbf{B}\mathbf{B}^\top\nabla^2\phi_{n_k}(x) =& -\sqrt{2\overline{\lambda}_k}\langle x, f_k\rangle\frac{\partial J_{n_{k1},n_{k2}}}{\partial z_k} \\ &- \sqrt{2}\lambda_k\langle x, \overline{f}_k\rangle\frac{\partial J_{n_{k1},n_{k2}}}{\partial \overline{z}_k} + 2\sigma_k^2\frac{\partial^2 J_{n_{k1},n_{k2}}}{\partial z\partial\overline{z}} \\ =& [-(n_{k1}+n_{k2})a_k + i(n_{k1}-n_{k2})b_k]J_{n_{k1},n_{k2}}. \end{aligned}$$

As for the cross terms, the normality of $\mathbf{A}$ implies that it is identically equal to zero. Therefore, we have $\mathcal{A}\phi_{\mathbf{n}}(x) = \mu_{\mathbf{n}}\phi_{\mathbf{n}}(x)$.

The above result also applies if $\mathbf{B}$ were a scalar multiple of an orthogonal matrix instead of being simultaneously diagonalizable with $\mathbf{A}$: i.e., when $\mathbf{B} = \sigma\mathbf{P}$ and $\mathbf{P}^\top\mathbf{P} = \mathbf{P}\mathbf{P}^\top = \mathbf{I}$.

A.3.4 Applications of the special case eigenfunctions

The eigenfunctions for the special cases above form complete orthonormal bases in $L^2(\nu)$, where ν is the invariant measure for the associated stochastic processes [24, 90]. The invariant density of ν is a normal distribution with mean zero and covariance $\Sigma = \int_0^T e^{s\mathbf{A}}\mathbf{B}\mathbf{B}^\top e^{s\mathbf{A}^\top}\mathrm{d}s$ [65]. Any function $g \in L^2(\nu)$ can then be expanded as an infinite sum of eigenfunctions, and the expansion coefficients can be expressed in terms of an integral with respect to the invariant measure:

$$g(x) = \sum_{\mathbf{n}} g_{\mathbf{n}}\phi_{\mathbf{n}}(x), \text{ where } g_{\mathbf{n}} = \int g(x)\phi_{\mathbf{n}}(x)\mathrm{d}\nu(x). \tag{A.14}$$

The eigenfunctions of the $L^2(\nu)$-adjoint of the OU operator can also be found explicitly in this case. The adjoint operator is the Fokker–Planck operator of the stochastic process [90]. The adjoint operator applied to a density $p \in L^2$ is

$$\mathcal{A}^* p(x) = -\sum_{i=1}^{d} \frac{\partial}{\partial x_i}[(\mathbf{A}x)_i p(x)] + \sum_{i,j=1}^{d} \frac{\partial^2}{\partial x_i \partial x_j}\mathbf{Q}_{ij}p(x). \tag{A.15}$$

The adjoint eigenfunctions are then $q_{\mathbf{n}}(x) = \phi_{\mathbf{n}}(x)p(x)$ with eigenvalue $\mu_{\mathbf{n}}$, where $p(x)$ is the invariant density. Solutions of the Kolmogorov backward equation (KBE) and Fokker–Planck equations can then also be expressed in terms of the eigenfunctions. For example, the KBE with terminal condition $g \in L^2(\nu)$:

$$\begin{cases} \dfrac{\partial \Phi(t,x)}{\partial t} + \mathcal{A}\Phi(t,x) & = 0 \\ \Phi(T,x) & = g(x) \end{cases}$$

has solution

$$\Phi(t,x) = \sum_{\mathbf{n}} g_{\mathbf{n}} e^{\mu_{\mathbf{n}}(T-t)}\phi_{\mathbf{n}}(x).$$

The solution of the Fokker–Planck equation can be obtained similarly.

A.4 Computation of general eigenfunctions

Here we turn to the case where we only assume $\mathbf{A}$ is diagonalizable. While in theory we know that the eigenfunctions can be expressed in closed form by polynomials, there is no simple way of expressing them in terms of classical orthogonal polynomials. Instead, we have found that a tractable approach for computing the eigenfunctions is to choose a basis of polynomials defined by the left eigenvectors of $\mathbf{A}$. Then, the action of the OU operator on the basis can be exactly represented by a matrix and the eigenfunctions are found by solving a matrix eigenvalue problem. We choose the basis $\{\psi_{\mathbf{n}}(x)\}_{\mathbf{n}\in\mathcal{I}}$, where the functions are defined in (A.4) and $\mathcal{I} \subset \mathbb{N}_0^d$ is some index set. This particular basis is chosen since its components are eigenfunctions of the first term of the OU operator. As we will see, this basis leads to a sparse matrix representation of the OU operator. Observe the following computation:

$$\mathcal{A}\psi_{\mathbf{n}}(x) = \langle x, \mathbf{A}^\top \nabla \psi_{\mathbf{n}} \rangle + \operatorname{Tr}\big[\mathbf{Q}\nabla^2\psi_{\mathbf{n}}\big] = \mu_{\mathbf{n}}\psi_{\mathbf{n}} + \operatorname{Tr}\big[\mathbf{Q}\nabla^2\psi_{\mathbf{n}}\big].$$

We have that the trace term is

$$\begin{aligned}
\operatorname{Tr}\big[\mathbf{Q}\nabla^2\psi_{\mathbf{n}}\big] &= \sum_{k=1}^{d} \operatorname{Tr}\big[\mathbf{Q}\nabla^2\psi_{n_k}(x)\big] \prod_{j=1, j\neq k}^{d} \psi_{n_j}(x) \\
&\quad + 2\sum_{k=1}^{d}\sum_{j=k+1}^{d} \operatorname{Tr}\Big[\mathbf{Q}\nabla\psi_{n_k}\nabla\psi_{n_j}^\top\Big] \prod_{l=1, l\neq k, l\neq j}^{d} \psi_{n_l} \\
&= \sum_{k=1}^{d} \operatorname{Tr}\big[\mathbf{Q}f_k f_k^\top\big] n_k(n_k-1)\langle x, f_k\rangle^{n_k-2} \prod_{j=1, j\neq k}^{d} \psi_{n_j}(x) \\
&\quad + 2\sum_{k=1}^{d}\sum_{j=k+1}^{d} \operatorname{Tr}\big[\mathbf{Q}f_k f_j^\top\big] n_k n_j \langle x, f_k\rangle^{n_k-1}\langle x, f_j\rangle^{n_j-1} \prod_{l=1, l\neq k, l\neq j}^{d} \psi_{n_l}.
\end{aligned}$$

In more compact notation, we write

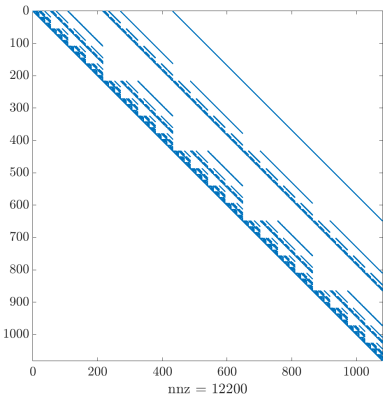
$$\begin{aligned}\mathcal{A}\psi_{\mathbf{n}}(x) =&\mu_{\mathbf{n}}\psi_{\mathbf{n}}(x) + \sum_{k=1}^{d}\langle \overline{f}_k, \mathbf{Q}f_k\rangle n_k(n_k-1)\psi_{\mathbf{m}^{(k)}}(x) \\ &+ 2\sum_{k=1}^{d}\sum_{j=k+1}^{d}\langle \overline{f}_j, \mathbf{Q}f_k\rangle n_k n_j \psi_{\mathbf{m}^{(kj)}}(x)\end{aligned} \tag{A.16}$$

where all entries of $\mathbf{m}^{(k)}$ and $\mathbf{m}^{(kj)}$ are equal to the corresponding entries of $\mathbf{n}$ except for $m_k^{(k)} = n_k - 2$, and $m_k^{(kj)} = n_k - 1$ and $m_j^{(kj)} = n_j - 1$. Therefore, as long as $\mathbf{m}^{(k)}$ and $\mathbf{m}^{(kj)}$ are in $\mathcal{I}$, then $\mathcal{A}\psi_{\mathbf{n}}(x)$ is contained in the span of $\{\psi_{\mathbf{n}}(x)\}_{\mathbf{n}\in\mathcal{I}}$. For practical computation, it is necessary to order the basis; lexicographical ordering is one obvious choice, but the choice is arbitrary and left to the user. Each basis function corresponds to an element of the standard basis, i.e., if there are $N = |\mathcal{I}|$ basis functions, then the k-th element of the basis corresponds to the vector in $\mathbb{R}^N$ with 1 in the k-th entry and zero everywhere else. The matrix representation of $\mathcal{A}$ is then $\mathbf{M} = [\mathcal{A}\psi_{\mathbf{n}_1} \cdots \mathcal{A}\psi_{\mathbf{n}_R}]$.

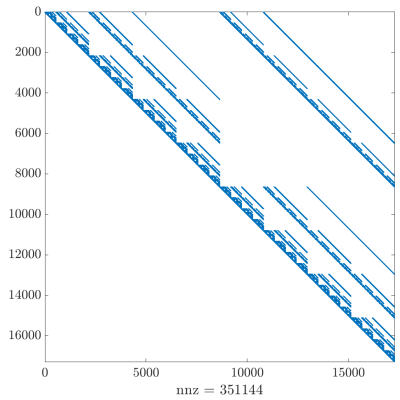
Suppose we are attempting to compute the eigenfunction with index $\mathbf{n}$. Based on (B.1), since the OU operator is a differential operator, $\mathcal{A}\psi_{\mathbf{n}}$ is itself a polynomial with index *less than* $\mathbf{n}$ in the lexicographical ordering. This would require at most $N = \prod_{k=1}^{d}(n_k+1)$ basis functions to span all the polynomials up to and including multi-index $\mathbf{n}$. The resulting matrix representation of $\mathbf{M}$ would then be an $N \times N$ matrix. However, (B.1) implies that $\mathcal{A}\psi_{\mathbf{n}}$ is dependent on at most $\frac{1}{2}(d^2+d+2)$ terms, which does not grow with the number of basis functions. Therefore, the resulting matrix is often quite sparse when many basis functions are considered. Solving the matrix eigenvalue problem would give *all* of the eigenfunctions of $\mathcal{A}$ with index up to and including $\mathbf{n}$.

Furthermore, if one only wishes to compute a single eigenfunction corresponding to index $\mathbf{n}$ (rather than all the eigenfunctions with total degree less than or equal to $\mathbf{n}$), then one does not need to include all the basis functions with index less than or equal to $\mathbf{n}$. For example, when $d = 2$ and we wish to compute the eigenfunction with index $(2,3)$, then the basis functions needed to express this eigenfunction have indices $\{(2,3),(2,1),(1,2),(1,0),(0,3),(0,1)\}$. In Figure A-1, we show the sparsity pattern of

two matrix representations of the OU operator in high dimensions. Lexicographical ordering was used in constructing these matrices. Notice that in Figure A-1a, the matrix has size 1080×1080; in contrast, the matrix would be of size 2160×2160 if all indices less than or equal to $\mathbf{n}$ were included in the basis. Similarly, in Figure A-1b, the matrix has size 17280×17280 rather than 34560×34560. The matrices were constructed by brute force, but they exhibit an interesting sparsity structure: for example, in Figure A-1b, only 0.12% of the matrix entries are nonzero. In future work, it may be interesting to investigate computationally efficient and structure-exploiting techniques for automatically constructing these matrices. As explored in [71], the



(a) $d = 6$, $l' = 4$, $l = 5$, $\mathbf{n} = (4, 3, 2, 2, 2, 3)$. The matrix is of size 1080×1080.



(b) $d = 9$, $l' = 5$, $l = 7$, $\mathbf{n} = (1, 3, 3, 2, 2, 1, 3, 4, 2)$. The matrix is of size 17280×17280.

Figure A-1: Sparsity patterns for two different matrix representations of $\mathcal{A}$.

eigenfunctions can be computed recursively via ladder operators. One could, therefore, express the ladder operators in terms of the basis we have chosen here, so that other eigenfunctions can be generated (given some initial high order eigenfunction).

Lastly, we comment on numerical methods for solving this matrix eigenvalue problem. Recall that given an index $\mathbf{n}$, the corresponding eigenvalue $\mu_{\mathbf{n}}$ is known exactly by Proposition 9, which means that only the eigenvectors need to be found. This means that only the nullspace of $\mathbf{M} - \mu_{\mathbf{n}}\mathbf{I}$ needs to be computed. In addition, $\mathbf{M} - \mu_{\mathbf{n}}\mathbf{I}$ is an *upper triangular* matrix, which means that if $\mathbf{M}$ can be stored in memory (even in a sparse fashion), then the reduced row echelon form of the matrix

can be easily computed and the nullspace can be found trivially. If only matrix-vector multiplies $\mathbf{M}v$ are accessible, the Arnoldi iteration can be employed to find the eigenvectors iteratively [116].

Remark 2. *One may ask if there is another choice of basis such that the number of terms produced by the trace term can be reduced. For example, a tempting choice is to use the basis defined in* (A.13)*. We found that this choice yields a more complicated expression that is similar to* (B.1) *without making the resulting matrix representation sparser.*

Remark 3. *Our approach is similar to that of [93], which computes eigenfunctions of the OU operator in the case that* $\mathbf{A}$ *is* not *diagonalizable (in contrast with the present setting). More specifically, [93] fixes a basis of polynomials (in fact, the tensorized Hermite polynomials) and seeks a finite-dimensional representation of the OU operator in that basis. However, eigenvalue problems of more than* $d = 3$ *dimensions were not studied.*

A.5 Discussion

We have presented a new approach for computing eigenfunctions of Ornstein–Uhlenbeck operators, in a general setting where the matrix $\mathbf{A}$ is diagonalizable. We first collect results for special cases, e.g., when $\mathbf{A}$ is self-adjoint or normal, and write explicit expresssions for the eigenfunctions in terms of certain orthogonal polynomials. We then address the general setting, where we show that by using a judicious choice of basis, one can compute eigenfunctions of any order, and in arbitrary dimension, by solving a sparse eigenvalue problem. The resulting matrix representation of the OU operator exhibits interesting structure that can be exploited to solve the associated eigenvalue problem efficiently.

These eigenfunctions have been found to be useful for applications such as simulating rare events [137] and approximating solutions to the Fokker–Planck equation [71]. We anticipate that this approach will be relevant for many other applications.

Appendix B

Supplementary Material for Chapter 5: The effects of discretization for irreversible Langevin dynamics

In this section, we study the effects of discretization in the setting of an irreversibly perturbed Langevin system. Results in full generality are, as yet, elusive; therefore we only consider a Gaussian example, as it still provides insight into how irreversibility interacts with discretization in impacting the asymptotic and finite sample bias and variance of the long term average estimator. While we do not present the results when a stochastic gradient is used, we note that the results are similar and can be easily extended based on what we present here. Recall that,

$$\mathbf{A} = \frac{1}{2}(\mathbf{I}+\mathbf{J})(\mathbf{\Gamma}_\theta + N\mathbf{\Gamma}_X), \quad \mathbf{D} = \frac{1}{2}(\mathbf{I}+\mathbf{J})\left(\mathbf{\Gamma}_X \sum_{i=1}^{N} X_i\right), \text{ where } \mathbf{J} = \delta \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

For this analysis, all precision matrices are 2×2 scalar matrices. That is, we assume $\mathbf{\Gamma}_\theta = \sigma_\theta^{-2}\mathbf{I}$, $\mathbf{\Gamma}_X = \sigma_X^{-2}\mathbf{I}$. This is distinct from the example in Section 5.3.1, since the precision matrices there are diagonal but not scalar. Let $b = \frac{1}{2\sigma_X^2}$ and $S_X = \sum_{k=1}^{N} X_i$, so that $\mathbf{D} = b(\mathbf{I}+\mathbf{J})S_X$.

We summarize our findings here. For fixed discretization size h and scalar precision

matrices as defined above, and introducing the irreversible perturbation scaled by δ, we find the following:

- The asymptotic bias for linear observables is zero, that is, $\mathbb{E}[\theta_\infty] = \mu_p$;
- The asymptotic variance for linear observables increases. We found that

$$\operatorname{Tr} \operatorname{Var}[\theta_\infty] = \frac{2}{2a - ha^2(1+\delta^2)} \tag{B.1}$$

  where $a = 0.5(1/\sigma_\theta^2 + N/\sigma_X^2)$;
- The finite time estimator for the observable $\phi(\theta) = \theta_1 + \theta_2$ has lower bias and variance;
- The finite time estimator for the observable $\phi(\theta) = \|\theta\|^2$ has higher bias and variance.

We focus on the finite time results and omit the asymptotic results, since the the former case is of more practical interest. The computations related to both are similar.

Finite time analysis: bias for linear observables. We study how the magnitude of the irreversibility, characterized by δ, impacts the mean-squared error $\text{MSE} = \mathbb{E}\big[\|\bar{\theta}_K - \mu_p\|^2\big]$ where $\bar{\theta}_K = \frac{1}{K}\sum_{k=0}^{K-1}\theta_k$. We approach this quantity via its bias-variance decomposition:

$$\text{MSE} = \big\|\mathbb{E}[\bar{\theta}_K] - \mu_p\big\|^2 + \operatorname{Tr} \operatorname{Var}\big(\bar{\theta}_K\big). \tag{B.2}$$

First, we compute the expected value of the sample average $\mathbb{E}\big[\bar{\theta}_K\big] = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\theta_k]$. For simplicity, we assume that the initial condition is always $\theta_0 = \mathbf{0}$. For any k, we

have

$$\begin{aligned}\mathbb{E}[\theta_k] &= (\mathbf{I} - h\mathbf{A})\,\mathbb{E}[\theta_{k-1}] + h\mathbf{D} \\ &= (\mathbf{I} - h\mathbf{A})^k\theta_0 + h\sum_{n=0}^{k-1}(\mathbf{I} - h\mathbf{A})^n\mathbf{D} \\ &= h(\mathbf{A}h)^{-1}(\mathbf{I} - (\mathbf{I} - h\mathbf{A})^k)\mathbf{D} \\ &= \mathbf{A}^{-1}\mathbf{D} - \mathbf{A}^{-1}(I - h\mathbf{A})^k\mathbf{D}.\end{aligned}$$

This yields

$$\begin{aligned}\mathbb{E}[\bar{\theta}_K] &= \frac{1}{K}\sum_{k=0}^{K-1}\big(\mathbf{A}^{-1}\mathbf{D} - \mathbf{A}^{-1}(\mathbf{I} - h\mathbf{A})^k\mathbf{D}\big) \\ &= \mathbf{A}^{-1}\mathbf{D} - \frac{1}{K}\mathbf{A}^{-1}(\mathbf{A}h)^{-1}(\mathbf{I} - (\mathbf{I} - h\mathbf{A})^K)\mathbf{D}.\end{aligned}$$

Since $\mu_p = \mathbf{A}^{-1}\mathbf{D}$, the bias is

$$\text{bias} = -\frac{1}{Kh}\mathbf{A}^{-2}(\mathbf{I} - (\mathbf{I} - h\mathbf{A})^K)\mathbf{D}.$$

The norm of the bias can in fact be computed. Note that $\mathbf{A}^2 = (1+\delta^2)a^2\mathbf{I}$ and we have

$$\begin{aligned}\|\text{bias}\|^2 &= \frac{1}{K^2h^2}\mathbf{D}^T(\mathbf{I} - (\mathbf{I} - h\mathbf{A}^T)^K)\mathbf{A}^{-2T}\mathbf{A}^{-2}(\mathbf{I} - (\mathbf{I} - h\mathbf{A})^K)\mathbf{D} \\ &= \frac{1}{K^2h^2a^4(1+\delta^2)^2}\mathbf{D}^T(\mathbf{I} - (\mathbf{I} - h\mathbf{A}^T)^K)(\mathbf{I} - (\mathbf{I} - h\mathbf{A})^K)\mathbf{D} \\ &= \frac{b^2}{K^2h^2a^4(1+\delta^2)^2}S_X^T(\mathbf{I}+\mathbf{J})^T(\mathbf{I} - (\mathbf{I} - \mathbf{A}^Th)^K)(\mathbf{I} - (\mathbf{I} - h\mathbf{A})^K)(\mathbf{I}+\mathbf{J})S_X.\end{aligned}$$

The inner matrix can be computed. Since each matrix above is simultaneously diagonalizable, we only need to consider the eigenvalues of each of the above matrices. Note that $\mathbf{I}+\mathbf{J}$ is a normal matrix, so we may write the eigenvalue decomposition

$\mathbf{I}+\mathbf{J}=\mathbf{PDP}^*$, where * denotes conjugate transpose, $\mathbf{Q}=\mathrm{diag}(1+i\delta, 1-i\delta)$, and

$$\mathbf{P}=\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ i & -i\end{bmatrix}$$

is orthogonal. Now note that

$$\mathbf{I}-(\mathbf{I}-h\mathbf{A})^K=\mathbf{P}\begin{bmatrix}1-(1-ah(1+i\delta))^K & 0\\ 0 & 1-(1-ah(1-i\delta))^K\end{bmatrix}\mathbf{P}^*,$$

which implies

$$(\mathbf{I}-(\mathbf{I}-h\mathbf{A}^T)^K)(\mathbf{I}-(\mathbf{I}-h\mathbf{A})^K)=|1-(1-ah(1+i\delta)^K)|^2\mathbf{I}.$$

Using the fact that $(\mathbf{I}+\mathbf{J})^T(\mathbf{I}+\mathbf{J})=(1+\delta^2)I$, and we have the following

$$\|\text{bias}\|^2=\frac{b^2}{K^2h^2a^4(1+\delta^2)}|1-(1-a(1+i\delta)h)^K|^2\|S_X\|. \tag{B.3}$$

To simplify further, we write $1-a(1+i\delta)h=re^{i\theta}$ where $r^2=(1-ah)^2+\delta^2a^2h^2$, and $\tan\theta=\delta ah/(1-ah)$. Then we obtain

$$\begin{aligned}\|\text{bias}\|^2&=\frac{b^2}{K^2h^2a^4(1+\delta^2)}|1-r^Ke^{i\theta K}|^2\|S_X\|^2\\ &=\frac{b^2}{K^2h^2a^4(1+\delta^2)}(1+r^{2K}-2r^K\cos K\theta)\|S_X\|^2.\end{aligned}$$

We know that $r<1$, since otherwise, the numerical scheme would be unstable. It is easy to see that for large, but not infinite, K, the bias decays as $\mathcal{O}(1/(Kh\sqrt{1+\delta^2}))$, so the introduction of irreversibility decreases the constant in front of the expression and therefore slightly improves the convergence of the bias.

Finite time analysis: variance for linear observables. For simplicity, we assume $\theta_0 = 0$. We compute $\mathrm{Tr}\mathrm{Var}(\bar{\theta}_K)$. We begin with

$$\mathrm{Tr}\,\mathrm{Var}(\bar{\theta}_K) = \mathrm{Tr}\,\mathbb{E}[\bar{\theta}_K\bar{\theta}_K^T] - \mathrm{Tr}\,\mathbb{E}[\bar{\theta}_K]\,\mathbb{E}[\bar{\theta}_K]^T$$

and compute these terms separately. It is difficult to surmise a relationship between δ and $\mathrm{Tr}\,\mathrm{Var}(\bar{\theta}_K)$ even with exact formulas, so we appeal to plots of the expressions to see that the variance decreases with irreversibility. We computed $\mathbb{E}[\bar{\theta}_K]$ in the previous section.

With the observation that

$$\mathbf{A}^{-2}(\mathbf{I} - (\mathbf{I} - h\mathbf{A})^K)\,\mathbb{E}[\mathbf{D}] = \frac{b}{a^2}\mathbf{P}\mathbf{Q}'\mathbf{P}^*\,\mathbb{E}[S_X]$$

where

$$\mathbf{Q}' = \begin{bmatrix} \frac{1-(1-ah(1+i\delta))^K}{1+i\delta} & 0 \\ 0 & \frac{1-(1-ah(1-i\delta))^K}{1-i\delta}, \end{bmatrix}$$

and $\mathbf{P}$ is defined in the previous section. We compute that

$$\mathrm{Tr}\,\mathbb{E}[\bar{\theta}_K]\,\mathbb{E}[\bar{\theta}_K]^T = \|\mu_p\|^2 + \|\mathrm{bias}\|^2 - \frac{2b^2}{Kha^3(1+\delta^2)}\mathrm{Re}\{(1-i\delta)(1-ah(1+i\delta))^K\}\|\mathbb{E}[S_X]\|^2. \tag{B.4}$$

The other term is more complicated and needs to be approached more carefully. Observe that

$$\mathrm{Tr}\,\mathbb{E}[\bar{\theta}_K\bar{\theta}_K^T] = \frac{1}{K^2}\sum_{i,j=1}^{K}\mathrm{Tr}\,\mathbb{E}[\theta_i\theta_j^T] = \frac{1}{K^2}\left(\sum_{i=0}^{K-1}\mathrm{Tr}\,\mathbb{E}[\theta_i\theta_i^T] + 2\sum_{i<j=0}^{K-1}\mathrm{Tr}\,\mathbb{E}[\theta_i\theta_j^T]\right).$$

We take each term individually. To compute $\mathbb{E}[\theta_k\theta_k^T]$, it is actually better to consider the covariance matrix of θ_k, $\Sigma_k = \mathbb{E}[\theta_k\theta_k^T] - \mathbb{E}[\theta_k]\,\mathbb{E}[\theta_k]^T$.

We first compute

$$\begin{aligned}
\mathbb{E}[\theta_k\theta_k^T] =&(\mathbf{I}-h\mathbf{A})\,\mathbb{E}[\theta_{k-1}\theta_{k-1}^T](\mathbf{I}-h\mathbf{A})^T + h^2\mathbf{D}\mathbf{D}^T + h\mathbf{I}\\
&+(\mathbf{I}-h\mathbf{A})\,\mathbb{E}[\theta_{k-1}]\mathbf{D}^T h + h\mathbf{D}\,\mathbb{E}[\theta_{k-1}^T](\mathbf{I}-h\mathbf{A})^T\\
\mathbb{E}[\theta_k]\,\mathbb{E}[\theta_k]^T =&(\mathbf{I}-h\mathbf{A})\,\mathbb{E}[\theta_{k-1}]\,\mathbb{E}[\theta_{k-1}]^T(\mathbf{I}-h\mathbf{A})^T + \mathbf{D}\,\mathbb{E}[\theta_{k-1}]^T(\mathbf{I}-h\mathbf{A})^T\\
&+(\mathbf{I}-h\mathbf{A})\,\mathbb{E}[\theta_{k-1}]\mathbf{D} + h^2\mathbf{D}\mathbf{D}^T
\end{aligned}$$

which imply the following recurrence relation. Assuming $\Sigma_0 = \mathbf{0}$, we have

$$\begin{aligned}
\Sigma_k = (\mathbf{I}-h\mathbf{A})\Sigma_{k-1}(\mathbf{I}-h\mathbf{A})^T + h\mathbf{I} &= h\sum_{n=0}^{k-1}\big((\mathbf{I}-h\mathbf{A})(\mathbf{I}-h\mathbf{A})^T\big)^n\\
&= h\sum_{n=0}^{k-1}(\mathbf{I}-(\mathbf{A}+\mathbf{A}^T)h + h^2\mathbf{A}\mathbf{A}^T)^n\\
&= ((\mathbf{A}+\mathbf{A}^T) - h^2\mathbf{A}\mathbf{A}^T)^{-1}(\mathbf{I}-(\mathbf{I}-(\mathbf{A}+\mathbf{A}^T)h + \mathbf{A}\mathbf{A}^T h^2)^k).
\end{aligned}$$

Let $s = 1-2ah+h^2a^2(1+\delta^2)$, then, by recalling that $\mathbf{A}+\mathbf{A}^T = 2a\mathbf{I}$ and $\mathbf{A}\mathbf{A}^T = a^2(1+\delta^2)\mathbf{I}$, the above sum is equal to $\frac{1-s^k}{1-s}h\mathbf{I}$. Therefore,

$$\operatorname{Tr}\Sigma_k = \frac{2h(1-s^k)}{1-s}. \tag{B.5}$$

Meanwhile note that

$$\mathbb{E}[\theta_k] = \mu_p - \mathbf{A}^{-1}(\mathbf{I}-h\mathbf{A})^k\mathbf{D}.$$

Therefore,

$$\begin{aligned}
\operatorname{Tr}\mathbb{E}[\theta_k]\,\mathbb{E}[\theta_k]^T = \mathbb{E}[\theta_k]^T\,\mathbb{E}[\theta_k] =&\|\mu_p\|^2 + \mathbf{D}^T(\mathbf{I}-h\mathbf{A}^T)^k\mathbf{A}^{-T}\mathbf{A}^{-1}(\mathbf{I}-h\mathbf{A})^k\mathbf{D}\\
&-2\mu_p^T\mathbf{A}^{-1}(\mathbf{I}-h\mathbf{A})^k\mathbf{D}\\
=&\|\mu_p\|^2 + \frac{s^k b^2}{a^2(1+\delta^2)}\|S_X\|^2 - 2\mu_p^T\mathbf{A}^{-1}(\mathbf{I}-h\mathbf{A})^k\mathbf{D}.
\end{aligned}$$

We now take the sum for each expression from $k = 0$ to $K - 1$. We have

$$\begin{aligned}\sum_{i=0}^{K-1} \operatorname{Tr} \Sigma_i &= 2h\left(\frac{K}{1-s} - \sum_{i=0}^{K-1} \frac{s^i}{1-s}\right) \\ &= 2h\left(\frac{K}{1-s} - \frac{1-s^K}{(1-s)^2}\right)\end{aligned}$$

and

$$\begin{aligned}\sum_{i=0}^{K-1} \mathbb{E}[\theta_i]^T \, \mathbb{E}[\theta_i] &= K\|\mu_p\|^2 + \frac{(1-s^K)b^2}{a^2(1+\delta^2)(1-s)}\|S_X\|^2 - 2\mu_p^T \mathbf{A}^{-1}(h\mathbf{A})^{-1}(\mathbf{I} - (\mathbf{I} - h\mathbf{A})^K)\mathbf{D} \\ &= K\|\mu_p\|^2 + \frac{(1-s^K)b^2}{a^2(1+\delta^2)(1-s)}\|S_X\|^2 - 2\mu_p^T h^{-1}\mathbf{A}^{-2}(\mathbf{I} - (\mathbf{I} - h\mathbf{A})^K)\mathbf{D}.\end{aligned}$$

For the cross-terms, observe that we may write

$$\sum_{i<j=0}^{K-1} \theta_i \theta_j^T = \sum_{i=0}^{K-1} \theta_i \sum_{j=i+1}^{K-1} \theta_j^T \tag{B.6}$$

which can be simplified further. First note that

$$\begin{aligned}\theta_j &= (\mathbf{I} - h\mathbf{A})\theta_{j-1} + \mathbf{D}h + \sqrt{h}\xi_{j-1} \\ &= (\mathbf{I} - h\mathbf{A})^{j-i}\theta_i + h\sum_{n=0}^{j-1-i}(\mathbf{I} - h\mathbf{A})^n \mathbf{D} + \sqrt{h}\sum_{n=0}^{j-i-1}(\mathbf{I} - h\mathbf{A})^n \xi_{j-1-n}.\end{aligned}$$

Plugging this expression into the double sum above, we have

$$\sum_{i=0}^{K-1} \theta_i \sum_{j=i+1}^{K-1} \left[\theta_i^T (\mathbf{I} - \mathbf{A}^T h)^{j-i} + h\sum_{n=0}^{j-1-i} \mathbf{D}^T (\mathbf{I} - \mathbf{A}^T h)^n + \sqrt{h}\sum_{n=0}^{j-i-1} \xi_{j-1-n}^T (\mathbf{I} - \mathbf{A}^T h)^n\right]. \tag{B.7}$$

Taking expectations, we have

$$\sum_{i=0}^{K-1}\sum_{j=i+1}^{K-1}\left[\mathbb{E}[\theta_i\theta_i^T](\mathbf{I}-h\mathbf{A}^T)^{j-i}+h\,\mathbb{E}[\theta_i]\mathbf{D}^T\sum_{n=0}^{j-1-i}(\mathbf{I}-h\mathbf{A}^T)^n\right]$$
$$=\sum_{i=0}^{K-1}\sum_{j=i+1}^{K-1}\left[\mathbb{E}[\theta_i\theta_i^T](\mathbf{I}-h\mathbf{A}^T)^{j-i}+h\,\mathbb{E}[\theta_i]\mathbf{D}^T(\mathbf{A}^Th)^{-1}(\mathbf{I}-(\mathbf{I}-\mathbf{A}^Th)^{j-i})\right].$$

Carrying out the computation for the first term, we have

$$F=\sum_{i=0}^{K-1}\mathbb{E}[\theta_i\theta_i^T]\sum_{j=i+1}^{K-1}(\mathbf{I}-\mathbf{A}^Th)^{j-i}$$
$$=\sum_{i=0}^{K-1}\mathbb{E}[\theta_i\theta_i^T](\mathbf{I}-\mathbf{A}^Th)(\mathbf{A}^Th)^{-1}(\mathbf{I}-(\mathbf{I}-\mathbf{A}^Th)^{K-1-i}).$$

For the second term we have,

$$\sum_{i=0}^{K-1}\mathbb{E}[\theta_i]\mu_p^T\sum_{j=i+1}^{K-1}(\mathbf{I}-(\mathbf{I}-\mathbf{A}^Th)^{j-i})$$
$$=\sum_{i=0}^{K-1}\mathbb{E}[\theta_i]\mu_p^T\left[(K-1-i)\mathbf{I}-(\mathbf{I}-\mathbf{A}^Th)(\mathbf{A}^Th)^{-1}(\mathbf{I}-(\mathbf{I}-\mathbf{A}^Th)^{K-1-i})\right].$$

The summations are difficult to compute precisely, so we compute them by direct evaluation instead. For simplicity, we assume that $\mu_p=\mathbb{E}[S_X]=[0,0]^T$, σ_X and σ_θ are chosen such that $a=1$. For this scenario, the bias is zero and only the variance contributes to the MSE. The variance is

$$\mathrm{Tr}[\mathrm{Var}\,\overline{\theta}_K]=\frac{1}{K^2}\left(2h\left(\frac{K}{1-s}-\frac{1-s^K}{(1-s)^2}\right)+2\,\mathrm{Tr}\,F\right)$$

where $\mathbb{E}[\theta_i\theta_i^T]=\Sigma_i=\frac{1-s^i}{1-s}h\mathbf{I}$.

In Figure B-1 we plot the variance for varying choices of δ. In this plots, $h=0.001$, $K=2\times10^5$, and δ varies between zero and ten. We can clearly see that strengthening the irreversible perturbation leads to improvement of the squared bias and variance of the long term average estimator.

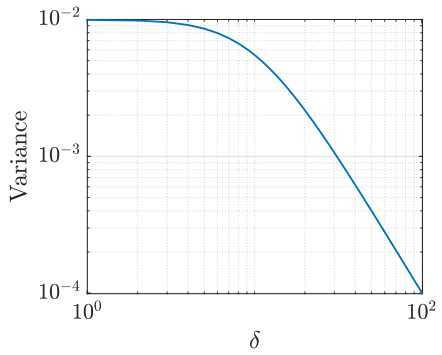


Figure B-1: Variance for different δ, fixed h.

Finite sample analysis for the quadratic observable $\phi(\theta) = \|\theta\|^2$. The previous finite sample results for the observable $\phi(\theta) = \theta_1 + \theta_2$ suggests that both the bias and variance of the long term average estimator goes down with a larger irreversible term. In this section, we show that this is actually a special case, and that when the observable is not linear, then the bias and variance may increase. We analyze the bias and variance of the long term average estimator of the observable $\phi(\theta) = \|\theta\|^2$. Define

$$\overline{\phi} = \int \phi(\theta)\pi(\theta)\mathrm{d}\theta, \quad \overline{\phi}_K = \frac{1}{K}\sum_{k=0}^{K-1} \phi(\theta_k). \tag{B.8}$$

As before, we assume that $\mu_p = [0,0]^T$, $\mathbb{E}[S_X] = \mathbf{0}$, and σ_x and σ_θ are chosen such that $a = 1$. We compute $|\mathbb{E}\,\overline{\phi}_K - \overline{\phi}|^2$ and $\mathrm{Var}\,\overline{\phi}_k$ and see how they vary with δ. From previous computations, we can show that

$$\mathbb{E}\,\overline{\phi}_K = 2h\left(\frac{1}{1-s} - \frac{1-s^K}{K(1-s)^2}\right), \tag{B.9}$$

where $s = 1 - 2ah + a^2h^2(1+\delta^2)$. Given this, the only term left to compute is the variance of the second moment of this observable:

$$\begin{aligned}\mathbb{E}\Big[\big(\overline{\phi}_K\big)^2\Big] &= \frac{1}{K^2}\,\mathbb{E}\left[\left(\sum_{k=0}^{K-1}\theta_k^T\theta_k\right)^2\right] \\ &= \frac{1}{K^2}\sum_{k=0}^{K-1}\mathbb{E}\big[(\theta_k^T\theta_k)^2\big] + \frac{2}{K^2}\,\mathbb{E}\sum_{k=0}^{K-1}\sum_{l=k+1}^{K-1}(\theta_k^T\theta_k)(\theta_l^T\theta_l).\end{aligned} \tag{B.10}$$

To compute the first sum, consider the following:

$$\theta_k^T\theta_k = \theta_{k-1}^T(\mathbf{I}-h\mathbf{A})^T(\mathbf{I}-h\mathbf{A})\theta_{k-1} + h\xi_{k-1}^T\xi_{k-1} + 2\sqrt{h}\theta_{k-1}^T(\mathbf{I}-h\mathbf{A})^T\xi_{k-1}$$

and so we have

$$\begin{aligned}(\theta_k^T\theta_k)^2 =& s^2(\theta_{k-1}^T\theta_{k-1})^2 + h^2(\xi_{k-1}^T\xi_{k-1})^2 + 4h(\xi_{k-1}^T(\mathbf{I}-h\mathbf{A})\theta_{k-1})^2 + 2s\theta_{k-1}^T\theta_{k-1}h\xi_{k-1}^T\xi_{k-1}\\ &+ 4s\sqrt{h}(\theta_{k-1}^T\theta_{k-1})\theta_{k-1}^T(\mathbf{I}-h\mathbf{A})^T\xi_{k-1} + 4h^{3/2}(\xi_{k-1}^T\xi_{k-1})\theta_{k-1}^T(\mathbf{I}-h\mathbf{A})^T\xi_{k-1}.\end{aligned}$$

Taking the expectation, we have

$$\begin{aligned}\mathbb{E}[(\theta_k^T\theta_k)^2] =& s^2\,\mathbb{E}[(\theta_{k-1}^T\theta_{k-1})^2] + h^2\,\mathbb{E}[(\xi_{k-1}^T\xi_{k-1})^2] + 4h\,\mathbb{E}\big[(\xi_{k-1}^T(\mathbf{I}-h\mathbf{A})\theta_{k-1})^2\big]\\ &+ 2sh\,\mathbb{E}\big[(\theta_{k-1}^T\theta_{k-1})(\xi_{k-1}^T\xi_{k-1})\big].\end{aligned}$$

After simplifying, we arrive at the following recurrence relation:

$$\mathbb{E}\big[(\theta_k^T\theta_k)^2\big] = s^2\,\mathbb{E}\big[(\theta_{k-1}^T\theta_{k-1})^2\big] + 8h^2 + 8sh\,\mathbb{E}\big[\theta_{k-1}^T\theta_{k-1}\big]. \tag{B.11}$$

Let $\beta_k = \mathbb{E}[(\theta_k^T\theta_k)^2]$, $\zeta_k = 8sh\,\mathbb{E}[\theta_k^T\theta_k]$, and $\kappa = 8h^2$. We have the following recurrence, which we solve

$$\begin{aligned}\beta_k &= s^2\beta_{k-1} + \zeta_{k-1} + \kappa\\ &= s^{2k}\beta_0 + \sum_{n=0}^{k-1} s^{2n}\zeta_{k-n-1} + \sum_{n=0}^{k-1} s^{2n}\kappa.\end{aligned}$$

From previous for the term ζ_k, we have

$$\begin{aligned}\beta_k &= \sum_{n=0}^{k-1} s^{2n}8sh\cdot 2h\frac{1-s^{k-n-1}}{1-s} + \kappa\frac{1-s^{2k}}{1-s^2}\\ &= \frac{16sh^2}{1-s}\sum_{n=0}^{k-1}(s^{2n}-s^{k+n-1}) + \frac{8h^2(1-s^{2k})}{1-s^2}\\ &= \frac{16sh^2}{1-s}\left(\frac{1-s^{2k}}{1-s^2} - s^{k-1}\frac{1-s^k}{1-s}\right) + \frac{8h^2(1-s^{2k})}{1-s^2}.\end{aligned}$$

Next we compute the summation of the cross terms. Define R_k such that

$$\sum_{k=0}^{K-1} R_k = \sum_{k=0}^{K-1} \sum_{l=k+1}^{K-1} \mathbb{E}\big[(\theta_k^T \theta_k)(\theta_l^T \theta_l)\big]. \tag{B.12}$$

We write

$$\begin{aligned}
\theta_l^T \theta_l &= s\theta_{l-1}^T \theta_{l-1} + h\xi_{l-1}^T \xi_{l-1} + 2\sqrt{h}\xi_{l-1}^T (\mathbf{I} - h\mathbf{A})\theta_{l-1} \\
&= s^{l-k}\theta_k^T \theta_k + \sum_{n=0}^{l-k-1} hs^n \xi_{l-n-1}^T \xi_{l-n-1} + 2\sqrt{h}s^n \xi_{l-n-1}^T (\mathbf{I} - h\mathbf{A})\theta_{l-n-1}.
\end{aligned}$$

This implies that

$$\begin{aligned}
R_k &= \sum_{l=k+1}^{K-1} \left(s^{l-k}\, \mathbb{E}[(\theta_k^T \theta_k)^2] + \sum_{n=0}^{l-k-1} 2hs^n\, \mathbb{E}[\theta_k^T \theta_k] \right) \\
&= \sum_{l=k+1}^{K-1} \beta_k s^{l-k} + 2h\, \mathbb{E}[\theta_k^T \theta_k] \sum_{n=0}^{l-k-1} s^n \\
&= \sum_{l=k+1}^{K-1} \beta_k s^{l-k} + 2h\, \mathbb{E}[\theta_k^T \theta_k] \frac{1 - s^{l-k}}{1-s} \\
&= \beta_k \sum_{l=k+1}^{K-1} s^{l-k} + \frac{2h\, \mathbb{E}[\theta_k^T \theta_k]}{1-s} \sum_{l=k+1}^{K-1} 1 - s^{l-k} \\
&= \beta_k \frac{s - s^{K-k}}{1-s} + \frac{2h\, \mathbb{E}[\theta_k^T \theta_k]}{1-s} \left(K - 1 - k - \frac{s - s^{K-k}}{1-s} \right).
\end{aligned}$$

To summarize, we have

$$\mathbb{E}\Big[\big(\overline{\phi}_K\big)^2\Big] = \frac{1}{K^2} \sum_{k=0}^{K-1} (\beta_k + 2R_k), \tag{B.13}$$

the squared bias is $(\mathbb{E}\,\overline{\phi}_K - 1)^2$ and the variance is $\mathbb{E}[(\overline{\phi}_K)^2] - \mathbb{E}[\overline{\phi}_K]^2$. These expressions are not simplifiable easily, so we plot these expressions and study their trends. In Figure B-2, we plot the squared bias and variance for fixed h and K and varying δ. In these plots, $h = 0.001$, $K = 2 \times 10^5$, and δ varies between zero and ten. Notice that for these choices, both the squared bias and variance increases as δ grows, showing

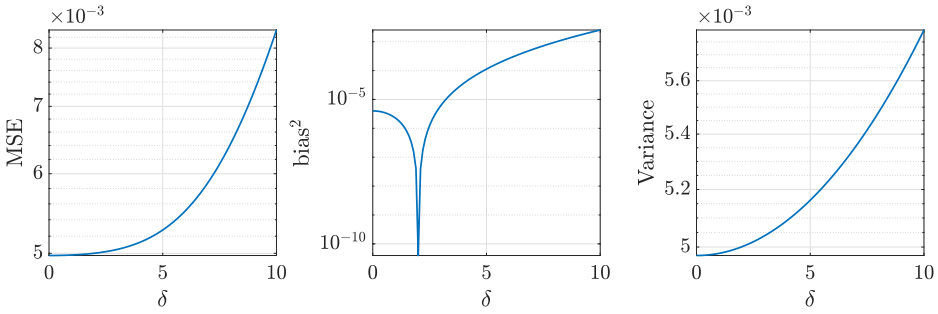


Figure B-2: Bias and variance of $\phi(\theta) = \|\theta\|^2$ for varying levels of irreversibility.

that for irreversibility provides no benefit, and in fact, harms the performance of the standard estimator.

Remark 4. *When discretization is considered, sampling properties in this simple example in which the drift is proportional to the state cannot be improved upon using irreversibility. Let us now explain this phenomenon from a theoretical point of view. In [97], the authors show that adding an irreversible perturbation to the generator of the diffusion process may decrease the spectral gap, and will never increase it. They further prove that in the continuous case, decreasing the spectral gap then decreases the asymptotic variance. However this improvement is not strict, that is, irreversibility is only guaranteed to not increase the spectral gap.*

Meanwhile, in [72], the authors consider irreversibility only in the context of linear systems, and rigorously study optimal irreversible perturbations that accelerate convergence to the invariant distribution. Their results show that when the drift matrix is proportional to the identity matrix, the spectral gap cannot be widened. Proposition 4 in [72] shows that the leading nonzero eigenvalue of the irreversibly perturbed drift matrix is bounded above by the leading nonzero eigenvalue of the original drift matrix and below by the trace of the original drift matrix over the dimension of the state space. The lower bound is then the optimal spectral gap. For a drift matrix that is a multiple of the identity, the upper and lower bounds are the same, which implies that the spectral gap can never decrease from its original value in the continuous case. After factoring in discretization, the irreversible perturbation increases stiffness of the system, which contributes to increased bias and variance in the resulting estimator.

Appendix C

Numerical solution to stochastic PDEs

We provide a brief review of methods for simulating stochastic PDEs. Much of this presentation is based on [62, 140]. We use these methods when discretizing the stochastic advection-diffusion equation in Chapter 2.4.2.

Stochastic PDEs are typically solved by formulating the equation as a stochastic differential equation on a Hilbert space of functions defined over some subset of $\mathbb{R}^d$. Let $H^2(D)$ be a Sobolev space over an open set $D \subset \mathbb{R}^d$. Let $\mathbf{A}$ be a compact self-adjoint linear operator that maps $H^2(D)$ to itself, and f be a possibly nonlinear function from $H^2(D)$ to itself. Stochastic PDEs are typically formulated in the following semilinear form,

$$\mathrm{d}X_t = [\mathbf{A}X_t + f(X_t)]\,\mathrm{d}t + \mathrm{d}W_t, \tag{C.1}$$

where f are nonlinear functions from $H^2(D)$ to itself, $X_t \in H^2(D)$ for all t, and W_t is an infinite-dimensional Wiener process. The inner product over this Sobolev space is

$$\langle \psi, \phi \rangle = \int_D \psi\phi \,\mathrm{d}x. \tag{C.2}$$

Theoretical details on infinite-dimensional Wiener processes can be found in [32].

C.1 Simulating infinite dimensional Wiener processes

Let H be a separable infinite dimensional Hilbert space and W_t be an H-valued Q-Wiener process. One may simulate this process using a series expansion. Let $\{e_k\}_{k\in\mathbb{N}}$ be an orthonormal basis of H comprised of eigenvectors of Q with eigenvalues $q_k > 0$. One can then represent W_t as follows,

$$W_t = \sum_{k=1}^{\infty} \sqrt{q_k}\beta_k(t)e_k, \tag{C.3}$$

where β_k are independent real-valued one-dimensional Wiener processes.

C.2 Exponential Euler schemes

The exponential Euler scheme is a type of Galerkin method in which the linear part of the projected SPDE is solved exactly. The nonlinear parts are integrated forwards in time using Duhamel's principle. We assume $\mathbf{A}$ admits an orthonormal basis $\{\phi_i\}_{i=1}^{\infty}$ in $L^2(D)$ with eigenvalues $-\lambda_i$ for $\lambda_i > 0$, where $\phi_k \in H^2(D) \cap H_0^1(D)$. Here, $H_0^1(D)$ denotes the Sobolev space whose functions satisfy Dirichlet boundary conditions.

Define a finite-dimensional subspace of $H^2(D)$ via a subset of the orthonormal basis $\{\phi_i\}_{i=1}^{N}$. We project the SPDE onto the resulting N-dimensional space and obtain a finite-dimensional representation of the SPDE,

$$\mathrm{d}X_t^N = \left[\mathbf{A}_N X_t^N + F_N(X_t^N)\right]\mathrm{d}t + \mathrm{d}W_t^N \tag{C.4}$$

where

$$\mathbf{A}_N v = \sum_{i=1}^{N} -\lambda_i \langle v, \phi_i\rangle\, \phi_i\,, \tag{C.5}$$

$$F_N = \mathcal{P}_N F\Big|_{\mathcal{X}_N} = \sum_{i=1}^{N} \langle f(X_t^N), \phi_i\rangle\, \phi_i. \tag{C.6}$$

The result of the projection is called the Itô-Galerkin stochastic ODE. The mild

representation of the solution is

$$X_t^N = e^{\mathbf{A}_N t} x_0^N + \int_0^t e^{\mathbf{A}_N (t-s)} F_N(X_s^N)\,\mathrm{d}s + \int_0^t e^{\mathbf{A}_N (t-s)}\,\mathrm{d}W_s^N. \tag{C.7}$$

Discretizing in time, we arrive at the following recurrence formula. Let $Y_k^N = X_{k\Delta t}^N$; then we have

$$Y_{k+1}^N = e^{\mathbf{A}_N \Delta t} Y_k^N + \mathbf{A}_N^{-1}\big(e^{\mathbf{A}_N \Delta t} - \mathbf{I}\big) F_N(Y_k^N) + \int_{t_k}^{t_{k+1}} e^{\mathbf{A}_N (t_{k+1}-s)}\,\mathrm{d}W_s. \tag{C.8}$$

This recurrence equation can be decoupled into N equations describing the evolutions of the coefficients of the solution. Let $Y_{k+1,i}^N$ and F_N^i denote the ith components of Y_{k+1}^N and F_N, respectively. Each component of Y_{k+1}^N evolves according to

$$Y_{k+1,i}^N = e^{-\lambda_i \Delta t} Y_{k,i}^N + \frac{1 - e^{-\lambda_i \Delta t}}{\lambda_i} F_N^i(Y_k^N) + \sqrt{q_k} \int_{t_k}^{t_{k+1}} e^{-\lambda_i (t_{k+1}-s)}\,\mathrm{d}\beta_i(s). \tag{C.9}$$

As for the last integral, note that any stochastic integral where the integrand is not dependent on the Brownian motion is Gaussian [85]. Furthermore, any Itô integral has mean zero (by virtue of being a martingale) and its variance can be computed via the Itô isometry. That is,

$$\mathbb{E}\left[\left(\int_{t_k}^{t_{k+1}} e^{-\lambda_i (t_{k+1}-s)}\,\mathrm{d}\beta_i(s)\right)^2\right] = \int_{t_k}^{t_{k+1}} e^{-2\lambda_i (s - t_{k+1})}\mathrm{d}s \tag{C.10}$$

$$= \frac{1}{2\lambda_i}\big(1 - e^{-2\lambda_i \Delta t}\big). \tag{C.11}$$

Therefore, the numerical algorithm for simulating the SPDE is

$$Y_{k+1,i}^N = e^{-\lambda_i \Delta t} Y_{k,i}^N + \frac{1 - e^{-\lambda_i \Delta t}}{\lambda_i} F_N^i(Y_k^N) + \sqrt{\frac{q_k}{2\lambda_i}(1 - e^{-2\lambda_i \Delta t})}\,\Delta W_k^i\,, \tag{C.12}$$

where $\Delta W_k^i \sim \mathcal{N}(0,1)$. When simulating the stochastic advection-diffusion equation, in particular, we have $\mathbf{A}v = \alpha v_{xx}$ and $f(v) = b v_x$.

Bibliography

[1] Shun-ichi Amari, Andrzej Cichocki, and Howard Hua Yang. A new learning algorithm for blind signal separation. In *Advances in neural information processing systems*, pages 757–763. Morgan Kaufmann Publishers, 1996.

[2] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

[3] Hassan Arbabi and Igor Mezić. Ergodic theory, dynamic mode decomposition, and computation of spectral properties of the Koopman operator. *SIAM Journal on Applied Dynamical Systems*, 16(4):2096–2126, 2017.

[4] Søren Asmussen and Peter W Glynn. *Stochastic simulation: algorithms and analysis*, volume 57. Springer Science & Business Media, 2007.

[5] Siu-Kui Au and James L Beck. Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic engineering mechanics*, 16(4):263–277, 2001.

[6] Ricardo Baptista, Olivier Zahm, and Youssef Marzouk. An adaptive transport framework for joint and conditional density estimation. *arXiv preprint arXiv:2009.10303*, 2020.

[7] Espen Bernton, Jeremy Heng, Arnaud Doucet, and Pierre E Jacob. Schrödinger bridge samplers. *arXiv preprint arXiv:1912.13170*, 2019.

[8] Tyrus Berry, Dimitrios Giannakis, and John Harlim. Nonparametric forecasting of low-dimensional dynamical systems. *Physical Review E*, 91(3):032915, 2015.

[9] Tyrus Berry and John Harlim. Nonparametric uncertainty quantification for stochastic gradient flows. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):484–508, 2015.

[10] Tyrus Berry and John Harlim. Variable bandwidth diffusion kernels. *Applied and Computational Harmonic Analysis*, 40(1):68–96, 2016.

[11] Vladimir Igorevich Bogachev. Ornstein–Uhlenbeck operators and semigroups. *Russian Mathematical Surveys*, 73(2):191, 2018.

[12] Zdravko I Botev and Dirk P Kroese. Efficient Monte Carlo simulation via the generalized splitting method. *Statistics and Computing*, 22(1):1–16, 2012.

[13] Michelle Boué and Paul Dupuis. A variational representation for certain functionals of Brownian motion. *The Annals of Probability*, 26(4):1641–1659, 1998.

[14] Charles-Edouard Bréhier and Tony Lelièvre. On a new class of score functions to estimate tail probabilities of some stochastic processes with adaptive multilevel splitting. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(3):033126, 2019.

[15] Nicolas Brosse, Eric Moulines, and Alain Durmus. The promises and pitfalls of stochastic gradient Langevin dynamics. *arXiv: 1811.10072*, 2018.

[16] Steven L Brunton and J Nathan Kutz. *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge University Press, 2019.

[17] Amarjit Budhiraja and Paul Dupuis. *Analysis and Approximation of Rare Events: Representations and Weak Convergence Methods*, volume 94. Springer, 2019.

[18] Marko Budišić, Ryan Mohr, and Igor Mezić. Applied Koopmanism. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 22(4):047510, 2012.

[19] Guillaume Carlier, Alfred Galichon, and Filippo Santambrogio. From Knothe's transport to Brenier's map and a continuation method for optimal transport. *SIAM Journal on Mathematical Analysis*, 41(6):2554–2576, 2010.

[20] Frédéric Cérou, Pierre Del Moral, Teddy Furon, and Arnaud Guyader. Sequential Monte Carlo for rare event estimation. *Statistics and computing*, 22(3):795–808, 2012.

[21] Frédéric Cérou, Bernard Delyon, Arnaud Guyader, and Mathias Rousset. On the asymptotic normality of adaptive multilevel splitting. *SIAM/ASA Journal on Uncertainty Quantification*, 7(1):1–30, 2019.

[22] Frédéric Cérou and Arnaud Guyader. Adaptive multilevel splitting for rare event analysis. *Stochastic Analysis and Applications*, 25(2):417–443, 2007.

[23] Frédéric Cérou, Arnaud Guyader, and Mathias Rousset. Adaptive multilevel splitting: Historical perspective and recent results. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(4):043108, 2019.

[24] Yong Chen, Yong Liu, et al. On the eigenfunctions of the complex Ornstein–Uhlenbeck operators. *Kyoto Journal of Mathematics*, 54(3):577–596, 2014.

[25] Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. Optimal steering of a linear stochastic system to a final probability distribution, Part I. *IEEE Transactions on Automatic Control*, 61(5):1158–1169, 2015.

[26] Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. Optimal steering of a linear stochastic system to a final probability distribution, Part III. *IEEE Transactions on Automatic Control*, 63(9):3112–3118, 2018.

[27] Eliodoro Chiavazzo, Roberto Covino, Ronald R Coifman, C William Gear, Anastasia S Georgiou, Gerhard Hummer, and Ioannis G Kevrekidis. Intrinsic map dynamics exploration for uncharted effective free-energy landscapes. *Proceedings of the National Academy of Sciences*, 114(28):E5494–E5503, 2017.

[28] Ronald R Coifman, Ioannis G Kevrekidis, Stéphane Lafon, Mauro Maggioni, and Boaz Nadler. Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. *Multiscale Modeling & Simulation*, 7(2):842–864, 2008.

[29] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

[30] Will Cousins, Miguel Onorato, Amin Chabchoub, and Themistoklis P Sapsis. Predicting ocean rogue waves from point measurements: An experimental study for unidirectional waves. *Physical Review E*, 99(3):032201, 2019.

[31] Nelida Črnjarić-Žic, Senka Maćešić, and Igor Mezić. Koopman operator spectrum for random dynamical systems. *Journal of Nonlinear Science*, pages 1–50, 2019.

[32] Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic equations in infinite dimensions*. Cambridge university press, 2014.

[33] Paolo Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1):313–329, 1991.

[34] Valentin De Bortoli, Arnaud Doucet, Jeremy Heng, and James Thornton. Simulating diffusion bridges with score matching. *arXiv preprint arXiv:2111.07243*, 2021.

[35] Thomas Dean and Paul Dupuis. Splitting for rare event simulation: A large deviation approach to design and analysis. *Stochastic processes and their applications*, 119(2):562–587, 2009.

[36] Giovanni Dematteis, Tobias Grafke, and Eric Vanden-Eijnden. Rogue waves and large deviations in deep sea. *Proceedings of the National Academy of Sciences*, 115(5):855–860, 2018.

[37] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Applications of mathematics. Springer, 1998.

[38] Joseph L Doob. *Classical potential theory and its probabilistic counterpart: Advanced problems*, volume 262. Springer Science & Business Media, 2012.

[39] A.B. Duncun, G.A. Pavliotis, and K.C. Zygalakis. Nonreversible langevin samplers: Splitting schemes, analysis and implementation. *arXiv preprint:1701.04247*, 2017.

[40] Paul Dupuis, Konstantinos Spiliopoulos, and Hui Wang. Importance sampling for multiscale diffusions. *Multiscale Modeling & Simulation*, 10(1):1–27, 2012.

[41] Paul Dupuis, Konstantinos Spiliopoulos, and Xiang Zhou. Escaping from an attractor: importance sampling and rest points I. *The Annals of Applied Probability*, 25(5):2909–2958, 2015.

[42] Paul Dupuis and Hui Wang. Importance sampling, large deviations, and differential games. *Stochastics: An International Journal of Probability and Stochastic Processes*, 76(6):481–508, 2004.

[43] Paul Dupuis and Hui Wang. Subsolutions of an Isaacs equation and efficient schemes for importance sampling. *Mathematics of Operations Research*, 32(3):723–757, 2007.

[44] Alain Durmus and Eric Moulines. High-dimensional bayesian inference via the unadjusted langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

[45] Lasse Ebener, Georgios Margazoglou, Jan Friedrich, Luca Biferale, and Rainer Grauer. Instanton based importance sampling for rare events in stochastic PDEs. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(6):063102, 2019.

[46] Paul Embrechts and Noël Veraverbeke. Estimates for the probability of ruin with special emphasis on the possibility of large claims. *Insurance: Mathematics and Economics*, 1(1):55–72, 1982.

[47] Samuel Feng, Philip Holmes, Alan Rorie, and William T Newsome. Can monkeys choose optimally when faced with noisy stimuli and unequal rewards? *PLoS computational biology*, 5(2):e1000284, 2009.

[48] Hans Föllmer. An entropy approach to the time reversal of diffusion processes. In *Stochastic Differential Systems Filtering and Control*, pages 156–163. Springer, 1985.

[49] Mark Iosifovich Freidlin and Alexander D Wentzell. Random perturbations. In *Random Perturbations of Dynamical Systems*, pages 15–43. Springer, 1998.

[50] Arnab Ganguly and P Sundar. Inhomogeneous functionals and approximations of invariant distribution of ergodic diffusions: Error analysis through central limit theorem and moderate deviation asymptotics. *Stochastic Processes and their Applications*, 133(C):74–110, 2021.

[51] Alfredo Garbuno-Inigo, Franca Hoffmann, Wuchen Li, and Andrew M Stuart. Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler. *SIAM Journal on Applied Dynamical Systems*, 19(1):412–441, 2020.

[52] Samuel Gershman, Matt Hoffman, and David Blei. Nonparametric variational inference. *arXiv preprint arXiv:1206.4665*, 2012.

[53] Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[54] Paul Glasserman, Yashan Wang, et al. Counterexamples in importance sampling for large deviations probabilities. *The Annals of Applied Probability*, 7(3):731–746, 1997.

[55] Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, pages 223–242, 2001.

[56] Carsten Hartmann, Omar Kebiri, Lara Neureither, and Lorenz Richter. Variational approach to rare event simulation using least-squares regression. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(6):063107, 2019.

[57] Carsten Hartmann, Lorenz Richter, Christof Schütte, and Wei Zhang. Variational characterization of free energy: Theory and algorithms. *Entropy*, 19(11):626, 2017.

[58] Matthew D Hoffman and Andrew Gelman. The no-U-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[59] Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. Mirrored langevin dynamics. *arXiv preprint arXiv:1802.10174*, 2018.

[60] Yuanhan Hu, Xiaoyu Wang, Xuefeng Gao, Mert Gürbüzbalaban, and Lingjiong Zhu. Non-convex optimization via non-reversible stochastic gradient Langevin dynamics. *arXiv preprint arXiv:2004.02823*, 2020.

[61] Chii-Ruey Hwang, Shu-Yin Hwang-Ma, and Shuenn-Jyi Sheu. Accelerating diffusions. *Annals of Applied Probability*, 15(2):1433–1444, 2005.

[62] Arnulf Jentzen and Peter E Kloeden. The numerical approximation of stochastic partial differential equations. *Milan Journal of Mathematics*, 77(1):205–244, 2009.

[63] Oliver Johnson. *Information theory and the central limit theorem*. World Scientific, 2004.

[64] Herman Kahn and Theodore E Harris. Estimation of particle transmission by random sampling. *National Bureau of Standards applied mathematics series*, 12:27–30, 1951.

[65] Ioannis Karatzas and Steven Shreve. *Brownian motion and stochastic calculus*, volume 113. Springer Science & Business Media, 2012.

[66] Omar Kebiri, Lara Neureither, and Carsten Hartmann. Adaptive importance sampling with forward-backward stochastic differential equations. *arXiv preprint arXiv:1802.04981*, 2018.

[67] Stefan Klus, Feliks Nüske, Sebastian Peitz, Jan-Hendrik Niemann, Cecilia Clementi, and Christof Schütte. Data-driven approximation of the Koopman generator: Model reduction, system identification, and control. *Physica D: Nonlinear Phenomena*, 406:132416, 2020.

[68] Bernard O Koopman. Hamiltonian systems and transformation in Hilbert space. *Proceedings of the National Academy of Sciences of the United States of America*, 17(5):315, 1931.

[69] Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[70] Todd K Leen and Robert Friel. Perturbation theory for stochastic learning dynamics. In *The 2011 International Joint Conference on Neural Networks*, pages 2031–2038. IEEE, 2011.

[71] Todd K Leen, Robert Friel, and David Nielsen. Eigenfunctions of the multi-dimensional linear noise Fokker-Planck operator via ladder operators. *arXiv preprint arXiv:1609.01194*, 2016.

[72] Tony Lelievre, Francis Nier, and Grigorios A Pavliotis. Optimal non-reversible linear drift for the convergence to equilibrium of a diffusion. *Journal of Statistical Physics*, 152(2):237–274, 2013.

[73] Christian Léonard. A survey of the schrödinger problem and some of its connections with optimal transport. *arXiv preprint arXiv:1308.0215*, 2013.

[74] Qianxiao Li, Felix Dietrich, Erik M Bollt, and Ioannis G Kevrekidis. Extended dynamic mode decomposition with dictionary learning: A data-driven adaptive spectral decomposition of the koopman operator. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(10):103111, 2017.

[75] Jingchen Liu, Jianfeng Lu, and Xiang Zhou. Efficient rare event simulation for failure problems in random media. *SIAM Journal on Scientific Computing*, 37(2):A609–A624, 2015.

[76] Samuel Livingstone and Mark Girolami. Information-geometric Markov chain Monte Carlo methods using diffusions. *Entropy*, 16(6):3074–3102, 2014.

[77] Jianfeng Lu and Konstantinos Spiliopoulos. Analysis of multiscale integrators for multiple attractors and irreversible langevin samplers. *Multiscale Modeling & Simulation*, 16(4):1859–1883, 2018.

[78] Alessandra Lunardi. On the Ornstein-Uhlenbeck operator in l^2 spaces with respect to invariant measures. *Transactions of the American Mathematical Society*, 349(1):155–169, 1997.

[79] G Margazoglou, L Biferale, R Grauer, K Jansen, D Mesterházy, T Rosenow, and R Tripiccione. Hybrid Monte Carlo algorithm for sampling rare events in space-time histories of stochastic fields. *Physical Review E*, 99(5):053303, 2019.

[80] Youssef Marzouk, Tarek Moselhy, Matthew Parno, and Alessio Spantini. An introduction to sampling via measure transport. *arXiv preprint arXiv:1602.05023*, 2016.

[81] A Mauroy, I Mezić, and Y Susuki. *Koopman Operator in Systems and Control*. Springer.

[82] Giorgio Metafune, Diego Pallara, and Enrico Priola. Spectrum of Ornstein-Uhlenbeck operators in L^p spaces with respect to invariant measures. *Journal of Functional Analysis*, 196(1):40–60, 2002.

[83] Igor Mezić. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, 41(1-3):309–325, 2005.

[84] Elisa Nicolato and Emmanouil Venardos. Option pricing in stochastic volatility models of the Ornstein-Uhlenbeck type. *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics*, 13(4):445–466, 2003.

[85] Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

[86] Michela Ottobre, Natesh S Pillai, and Konstantinos Spiliopoulos. Optimal scaling of the MALA algorithm with irreversible proposals for Gaussian targets. *Stochastics and Partial Differential Equations: Analysis and Computations*, pages 1–51, 2019.

[87] Art B. Owen. *Monte Carlo theory, methods and examples*. 2013.

[88] Iason Papaioannou, Wolfgang Betz, Kilian Zwirglmaier, and Daniel Straub. MCMC algorithms for subset simulation. *Probabilistic Engineering Mechanics*, 41:89–103, 2015.

[89] Matthew Parno and Youssef Marzouk. Transport map accelerated Markov chain Monte Carlo. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):645–682, 2018.

[90] Grigorios A Pavliotis. *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*, volume 60. Springer, 2014.

[91] Michele Pavon, Esteban G Tabak, and Giulio Trigila. The data-driven Schrödinger bridge. *arXiv preprint arXiv:1806.01364*, 2018.

[92] Baron Peters. *Reaction rate theory and rare events*. Elsevier, 2017.

[93] Yulei Rao, Jiying Wang, and Yong Chen. Jordan decomposition and geometric multiplicity for a class of non-symmetric Ornstein-Uhlenbeck operators. *Advances in Difference Equations*, 2014(1):1–14, 2014.

[94] Sebastian Reich. Data assimilation: The Schrödinger perspective. *Acta Numerica*, 28:635–711, 2019.

[95] Luc Rey-Bellet and Konstantinos Spiliopoulos. Irreversible Langevin samplers and variance reduction: a large deviations approach. *Nonlinearity*, 28(7):2081, 2015.

[96] Luc Rey-Bellet and Konstantinos Spiliopoulos. Variance reduction for irreversible langevin samplers and diffusion on graphs. *Electronic Communications in Probability*, 20, 2015.

[97] Luc Rey-Bellet and Konstantinos Spiliopoulos. Improving the convergence of reversible samplers. *Journal of Statistical Physics*, 164(3):472–494, 2016.

[98] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

[99] Luigi M Ricciardi and Laura Sacerdote. The Ornstein-Uhlenbeck process as a model for neuronal activity. *Biological cybernetics*, 35(1):1–9, 1979.

[100] Gareth O Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[101] L Chris G Rogers and David Williams. *Diffusions, Markov processes and martingales: Volume 2, Itô calculus*, volume 2. Cambridge University Press, 2000.

[102] Andreas Rößler. Runge–kutta methods for the strong approximation of solutions of stochastic differential equations. *SIAM Journal on Numerical Analysis*, 48(3):922–952, 2010.

[103] Clarence W Rowley, Igor Mezić, Shervin Bagheri, Philipp Schlatter, and Dan S Henningson. Spectral analysis of nonlinear flows. *Journal of Fluid Mechanics*, 641:115–127, 2009.

[104] Michael Salins and Konstantinos Spiliopoulos. Rare event simulation via importance sampling for linear SPDEs. *Stochastics and Partial Differential Equations: Analysis and Computations*, pages 1–39, 2017.

[105] Simo Särkkä and Arno Solin. *Applied stochastic differential equations*, volume 10. Cambridge University Press, 2019.

[106] Peter J Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of fluid mechanics*, 656:5–28, 2010.

[107] Erwin Schrödinger. *Über die umkehrung der naturgesetze*. Verlag der Akademie der Wissenschaften in Kommission bei Walter De Gruyter. Company, 1931.

[108] Jie Shen and Li-Lian Wang. Sparse spectral approximations of high-dimensional problems based on hyperbolic cross. *SIAM Journal on Numerical Analysis*, 48(3):1087–1109, 2010.

[109] Alessio Spantini, Daniele Bigoni, and Youssef Marzouk. Inference via low-dimensional couplings. *The Journal of Machine Learning Research*, 19(1):2639–2709, 2018.

[110] Konstantinos Spiliopoulos. Nonasymptotic performance analysis of importance sampling schemes for small noise diffusions. *Journal of Applied Probability*, 52(3):797–810, 2015.

[111] Steven H Strogatz. *Nonlinear dynamics and chaos with student solutions manual: With applications to physics, biology, chemistry, and engineering*. CRC press, 2018.

[112] Daniel W Stroock. *Elements of Stochastic Calculus and Analysis*. Springer.

[113] Naoya Takeishi, Yoshinobu Kawahara, and Takehisa Yairi. Learning koopman invariant subspaces for dynamic mode decomposition. In *NIPS*, 2017.

[114] Yee Whye Teh, Alexandre H Thiery, and Sebastian J Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *Journal of Machine Learning Research*, 17, 2016.

[115] Erik H Thiede, Dimitrios Giannakis, Aaron R Dinner, and Jonathan Weare. Galerkin approximation of dynamical quantities using trajectory data. *The Journal of chemical physics*, 150(24):244111, 2019.

[116] Lloyd N Trefethen and David Bau III. *Numerical linear algebra*, volume 50. SIAM, 1997.

[117] Nicolas Garcia Trillos and Daniel Sanz-Alonso. The Bayesian update: variational formulations and gradient flows. *Bayesian Analysis*, 15(1):29–56, 2020.

[118] Jonathan H Tu, Clarence W Rowley, Dirk M Luchtenburg, Steven L Brunton, and J Nathan Kutz. On dynamic mode decomposition: theory and applications. *Journal of Computational Dynamics*, 1(2), 2014.

[119] Belinda Tzen and Maxim Raginsky. Neural stochastic differential equations: Deep latent Gaussian models in the diffusion limit. *arXiv preprint arXiv:1905.09883*, 2019.

[120] Belinda Tzen and Maxim Raginsky. Theoretical guarantees for sampling and inference in generative models with latent diffusions. *arXiv preprint arXiv:1903.01608*, 2019.

[121] Eric Vanden-Eijnden. Transition path theory. In *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology Volume 1*, pages 453–493. Springer, 2006.

[122] Eric Vanden-Eijnden and Jonathan Weare. Rare event simulation of small noise diffusions. *Communications on Pure and Applied Mathematics*, 65(12):1770–1803, 2012.

[123] SR Srinivasa Varadhan. *Large deviations and applications*. SIAM, 1984.

[124] Francisco Vargas, Andrius Ovsianas, David Fernandes, Mark Girolami, Neil Lawrence, and Nikolas Nüsken. Bayesian learning via neural schr\" odinger-f\" ollmer flows. *arXiv preprint arXiv:2111.10510*, 2021.

[125] Oldrich Vasicek. An equilibrium characterization of the term structure. *Journal of financial economics*, 5(2):177–188, 1977.

[126] Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32:8094–8106, 2019.

[127] Manuel Villen-Altamirano, Jose Villen-Altamirano, et al. RESTART: A method for accelerating rare event simulations. *Queueing, performance and Control in ATM*, pages 71–76, 1991.

[128] Sebastian J Vollmer, Konstantinos C Zygalakis, and Yee Whye Teh. Exploration of the (non-) asymptotic bias and variance of stochastic gradient Langevin dynamics. *The Journal of Machine Learning Research*, 17(1):5504–5548, 2016.

[129] Mathias Wanner and Igor Mezić. Robust approximation of the stochastic Koopman operator. *arXiv preprint arXiv:2011.00078*, 2021.

[130] E Weinan, Weiqing Ren, and Eric Vanden-Eijnden. String method for the study of rare events. *Physical Review B*, 66(5):052301, 2002.

[131] E Weinan and Xiang Zhou. The gentlest ascent dynamics. *Nonlinearity*, 24(6):1831, 2011.

[132] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

[133] Matthew O Williams, Ioannis G Kevrekidis, and Clarence W Rowley. A data–driven approximation of the Koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015.

[134] Tatiana Xifara, Chris Sherlock, Samuel Livingstone, Simon Byrne, and Mark Girolami. Langevin diffusions and the Metropolis-adjusted langevin algorithm. *Statistics & Probability Letters*, 91:14–19, 2014.

[135] Benjamin Zhang, Youssef Marzouk, Byung-Young Min, and Tuhin Sahai. Rare event simulation of a rotorcraft system. In *2018 AIAA Non-Deterministic Approaches Conference*, page 1181, 2018.

[136] Benjamin Zhang, Youssef Marzouk, and Konstantinos Spiliopoulos. Geometry-informed irreversible perturbations for accelerated convergence of Langevin dynamics. *arXiv preprint arXiv:2108.08247*, 2021.

[137] Benjamin Zhang, Tuhin Sahai, and Youssef Marzouk. A Koopman framework for rare event simulation in stochastic differential equations. *arXiv preprint arXiv:2101.07330*, 2021.

[138] Benjamin J Zhang, Tuhin Sahai, and Youssef M Marzouk. Computing eigenfunctions of the multidimensional Ornstein-Uhlenbeck operator. *arXiv preprint arXiv:2110.09229*, 2021.

[139] Wei Zhang, Han Wang, Carsten Hartmann, Marcus Weber, and Christof Schütte. Applications of the cross-entropy method to importance sampling and optimal control of diffusions. *SIAM Journal on Scientific Computing*, 36(6):A2654–A2672, 2014.

[140] Zhongqiang Zhang and George Karniadakis. *Numerical methods for stochastic partial differential equations with white noise*, volume 196. Springer, 2017.