Sampling and Estimation on Manifolds using the Langevin Diffusion

N.B. When citing this work, cite the original published paper.

(article starts on next page)

Sampling and Estimation on Manifolds using the Langevin Diffusion

Karthik Bharath   Karthik.Bharath@nottingham.ac.uk
School of Mathematical Sciences
University of Nottingham
Nottingham, UK

Alexander Lewis   Alexander.Lewis@uni-goettingen.de
Institut für Mathematische Stochastik
Georg-August-Universität Göttingen
37077 Göttingen, Germany

Akash Sharma   akashs@chalmers.se
Department of Mathematical Sciences
Chalmers University of Technology and the University of Gothenburg
41296 Gothenburg, Sweden

Michael V. Tretyakov   Michael.Tretyakov@nottingham.ac.uk
School of Mathematical Sciences
University of Nottingham
Nottingham, UK

Editor: Anthony Lee

Abstract

Error bounds are derived for sampling and estimation using a discretization of an intrinsically defined Langevin diffusion with invariant measure $d\mu_\phi \propto e^{-\phi} dvol_g$ on a compact Riemannian manifold. Two estimators of linear functionals of μ_ϕ based on the discretized Markov process are considered: a time-averaging estimator based on a single trajectory and an ensemble-averaging estimator based on multiple independent trajectories. Imposing no restrictions beyond a nominal level of smoothness on ϕ, first-order error bounds, in discretization step size, on the bias and variance/mean-square error of both estimators are derived. The order of error matches the optimal rate in Euclidean and flat spaces, and leads to a first-order bound on distance between the invariant measure μ_ϕ and a stationary measure of the discretized Markov process. This order is preserved even upon using retractions when exponential maps are unavailable in closed form, thus enhancing practicality of the proposed algorithms. Generality of the proof techniques, which exploit links between two partial differential equations and the semigroup of operators corresponding to the Langevin diffusion, renders them amenable for the study of a more general class of sampling algorithms related to the Langevin diffusion. Conditions for extending analysis to the case of non-compact manifolds are discussed. Numerical illustrations with distributions, log-concave and otherwise, on the manifolds of positive and negative curvature elucidate on the derived bounds and demonstrate practical utility of the sampling algorithm.

Keywords: Stochastic differential equations on manifolds; Intrinsic Riemannian geometry; Weak approximation; Computing ergodic limits; Monte Carlo technique.

1. Introduction

The Langevin algorithm for sampling from a target probability measure $\mathrm{d}\mu_\phi \propto e^{-\phi(x)}\mathrm{d}x$ on $\mathbb{R}^q, q \geq 1$, is based on constructing a discrete time Markov process $\{X_n^h, n \geq 1\}$ with time-step size $h = T/N > 0$, $t_n = hn$, obtained by discretizing the ergodic Langevin diffusion

$$\mathrm{d}X(t) = -\frac{1}{2}\nabla\phi(X(t))\mathrm{d}t + \mathrm{d}B(t), \quad 0 \leq t \leq T, \tag{1.1}$$

where $B(t)$ is the q-dimensional standard Brownian motion. Popular amongst the numerical methods for constructing X_n^h is the (explicit) Euler method with Gaussian noise to approximate Brownian motion increments, under which various types of approximation errors in using X_n^h in place of $X(t_n)$ have been ascertained (e.g., Talay, 1990; Roberts and Tweedie, 1996; Dalalyan, 2017; Milstein and Tretyakov, 2007, 2021). The natural regime to examine error bounds for such algorithms is when step size $h \to 0$.

The Langevin diffusion (1.1) has been successfully used for sampling from posterior distributions in Bayesian inference (e.g., Roberts and Tweedie, 1996; Durmus and Moulines, 2017), for canonical-ensemble calculations in molecular dynamics (Leimkuhler and Matthews, 2015) and for approximating minimizers in convex/non-convex optimisation (e.g., Teh et al., 2016). Recently, there has been growing interest in finding effective ways to sample from distributions on manifolds, and in providing corresponding theoretical guarantees (e.g., Girolami and Calderhead, 2011; Barp et al., 2022; Gatmiry and Vempala, 2022; Cheng et al., 2022; Li and Erdogdu, 2023). Our work aims to provide theoretical guarantees for an intrinsically-defined Langevin algorithm for sampling from a target probability measure on Riemannian manifolds.

In a statistical context, when sampling and estimation are of primary interest, quantitative error bounds, in terms of the time-step h and integration time T, on two interconnected quantities related to distributional aspects of the approximations are relevant: bias of an estimator based on X_n^h of the linear functional $\mu_\phi(\varphi) := \int \varphi \mathrm{d}\mu_\phi$ over a class $\mathcal{H}$ of test functions φ; and an integral probability metric with respect to $\mathcal{H}$ between the stationary measure of X_n^h and μ_ϕ. Such bounds are referred to as *weak error bounds*. In statistical applications (e.g., Bayesian computing), the commonly used estimator of $\mu_\phi(\varphi)$ is based on a single trajectory of X_n^h until a large time $T = Nh$, known as the time-averaging estimator. Under suitable conditions on ϕ, when the law of $X(t)$ converges to its invariant measure μ_ϕ as $t \to \infty$, it is known that for the Euler method the optimal rate at which bias of the time-averaging estimator disappears is $O(h)$ (Talay, 1990; Milstein and Tretyakov, 2007; Mattingly et al., 2010; Milstein and Tretyakov, 2021); the first-order (in h) bound on the bias then results in a commensurate first-order bound on the integral probability metric.

Modern statistical and machine learning applications have generated a growing need for extending the above program to a q-dimensional smooth Riemannian manifold M. To better understand the challenges, some context on the necessary ingredients is helpful. Broadly, the type of application engenders two perspectives of, and hence coordinate systems on, M: (i) as a submanifold of $\mathbb{R}^k, k \geq q$, with embedded geometry inherited from that of $\mathbb{R}^k$; (ii) as a q-dimensional differentiable manifold equipped with an intrinsic geometry. These two points of view lead to different formulations of the Langevin diffusion on M with different invariant measures. The focus of this article is on the intrinsic perspective, and on investigating a geodesic-based sampling algorithm built on an intrinsic Langevin diffusion.

A Riemannian metric g prescribes a notion of curvature of the manifold M at a point and determines a unique intrinsic Brownian motion on it (e.g., Elworthy, 1982; Hsu, 2002; Wang, 2013). The curvature of M and behavior of the Brownian motion are intimately related: positive curvature makes curves come closer and thus prevents a Brownian motion from wandering away to infinity (non-explosion) while ensuring recurrence and fast convergence to ergodicity (e.g., Ichihara, 1982a,b); this is true even for its approximation based on a geodesic random walk (Mangoubi and Smith, 2018). Negative curvature, on the other hand, disperses curves and encourages Brownian motion to exit compact sets and drift away to infinity (explosion) while discouraging recurrence (e.g., Kendall, 1984). Ricci curvature at a point, which is, roughly, the average of the sectional curvatures of two-dimensional subspaces of the tangent space at the point, captures the notion of curvature alluded to above. Compactness (unboundedness) and positive (non-positive) curvature of M are closely related, although there are non-trivial exceptions depending on the Riemannian metric g (for instance, manifolds with hyperbolic metric g (Everitt and Maclachlan, 2000)). Thus, for negatively curved (typically) non-compact spaces, a uniform lower bound on the Ricci curvature is needed to ensure that the Brownian motion is well-behaved, while this is automatically true for positively curved (typically) compact manifolds. These issues, as one can now imagine, are subtler for a Brownian motion with drift.

It would thus be reasonable to expect the intrinsic curvature of M, and regularity and behavior at infinity of its drift, to influence error of approximations of an intrinsically defined Langevin diffusion and its invariant measure. This is apparently the case in recent works (e.g., Gatmiry and Vempala, 2022; Cheng et al., 2022), where curvature-dependent constant terms show up in upper bounds on distances between the stationary measure μ_ϕ^h of a discretized Markov processes X_n^h and μ_ϕ on M defined with respect to the Riemannian volume measure dvol_g. Importantly, however, the curvature-terms *do not* affect the claimed error order with respect to step size h.

A differentiable manifold M is locally Euclidean, covered by a family of charts (open subsets of $\mathbb{R}^q$) and smooth mappings between them. In a small region of M, local analysis of the weak error within a chart, where X solves the appropriate stochastic differential equation (SDE), proceeds as in $\mathbb{R}^q$ using the local coordinate representation of the metric g in the chart, which accounts for the local curvature. The challenge lies in adapting the analysis when one transitions to another chart, where the local representation of the metric g, and hence that of X, changes. Thus, on flat manifolds M with zero curvature, error analysis proceeds unabated as in $(\mathbb{R}^q, g)$. Indeed, Mattingly et al. (2010) demonstrate precisely this with weak error analysis when $M = \mathbb{T}^q \cong \mathbb{R}^q/\mathbb{Z}^q$, the q-dimensional compact torus, equipped with the flat (quotient) metric from $\mathbb{R}^q$. The main technical ingredient in their analysis was the link between the solution of a Poisson equation (given by a partial differential equation (PDE)) on $\mathbb{T}^q$ and the infinitesimal generator of the Langevin diffusion X. Aided by the fact that a weak error analysis merely requires bounding and matching of a handful of moments, assured by compactness of $\mathbb{T}^q$, an *optimal first-order* $O(h)$ weak error bound, matching the situation in $\mathbb{R}^q$, was established for general discretization schemes, including the Euler method.

It thus seems plausible to posit that weak error bounds of a suitable Euler method on compact manifolds M should match the optimal first-order rate seen on $\mathbb{T}^q$. This leads us to the two related questions of interest in this work, stated informally as:

(Q1) Can an intrinsic Euler discretization method of an intrinsically defined Langevin diffusion on a compact manifold M match the first-order weak error bounds obtained in the Euclidean setting?

(Q2) Is it possible to construct estimators of linear functionals of the target measure on M with first-order bias terms?

1.1 Contributions and related work

Our main contribution is in providing affirmative answers to (Q1) and (Q2), and in doing so demonstrate utility of the PDE-based tools in obtaining error bounds of Langevin-based sampling algorithms for compact manifolds.

Specifically, on a connected compact Riemannian manifold (M, g) without boundary, we consider an intrinsically defined Langevin diffusion

$$\mathrm{d}X(t) = -\frac{1}{2}\nabla\phi(X(t))\mathrm{d}t + \mathrm{d}B^M(t), \quad X(0) = x \in M,$$

with invariant measure $\mathrm{d}\mu_\phi \propto e^{-\phi(x)}\mathrm{dvol}_g$, where B^M is a Brownian motion on M, and its Euler discretization X_n^h based on moving along geodesics, or their approximations, determined by drift-dependent tangent direction and $\mathbb{R}^q$-valued noise. The resulting algorithm is termed *Riemannian Langevin*. The potential ϕ is not required to be (geodesically) convex. At every step X_n^h automatically belongs to the manifold M owing to use of the exponential map (or more generally, retractions), i.e., the algorithm respects the geometry of the problem. It is well-known in deterministic (Hairer et al., 2002) and stochastic (Leimkuhler and Matthews, 2015; Milstein and Tretyakov, 2021) numerical analysis that geometric integrators (i.e., numerical methods which naturally preserve geometric features of differential equations they approximate) are computationally more efficient (e.g., allowing larger time steps while still producing accurate results) for long time simulations as ones arising in sampling problems. The Riemannian Langevin algorithm considered here belongs to the class of such geometric integrators, and is hence suitable for sampling on manifolds.

Two estimators of the linear functional $\mu_\phi(\varphi)$ for a smooth φ are proposed: an *ensemble-averaging* estimator defined using multiple independent trajectories of X_n^h, and a *time-averaging* estimator based on a single trajectory. The former is of great practical merit. First, it enables decoupling and management of three sources of error: (i) length of the integration time T that controls proximity of X to its ergodic limit; (ii) discretization error controlled by the step size h; (iii) the number of independent trajectories of X_n^h that controls the Monte Carlo error (sample variance). Second, the estimator is particularly useful in a parallel computing environment (GPUs or array of CPUs) where independent trajectories of X_n^h can be generated in parallel and combined in the end. Analysis of the time-averaging estimator leads to first-order bound on a distance between the empirical measure μ_ϕ^h of X_n^h and μ_ϕ, with respect to a smooth class $\mathcal{H}$ of test functions.

Our technical analyses are based on exploiting links between the Markov semigroup of X and the Kolmogorov and Poisson PDEs; a positive answer to (Q1), bearing in mind the optimal error rate in $\mathbb{T}^q$, is realised by effectively 'flattening' the compact M, at least as far as the deviation of the stationary measure μ_ϕ^h of X_n^h from μ_ϕ is concerned. The price paid for the proof technique used is seen in the absence of *explicit* curvature-related

constants in the weak error bounds. However, in addition to deriving the optimal order for weak error, there are other important benefits to the semigroup approach which make it suitable for analyzing sampling algorithms: (i) analysis can be extended to sampling on non-compact manifolds (Section 5); (ii) variants of the Riemannian Langevin algorithm with many families of distributions for increments of the Markov chain X_n^h, and based on splitting methods, can be analysed in a unified manner (Section 3.3).

In existing works on SDE-based algorithms for sampling on manifolds (Gatmiry and Vempala, 2022; Cheng et al., 2022) weak convergence of the algorithm is established by first proving mean-square (strong) convergence, since weak convergence follows from mean-square; a consequence of this approach is usually a non-optimal weak convergence order when compared to the situation in $\mathbb{R}^q$. A notable exception to this is Grorud and Talay (1996), which we discuss in Section 3.1. Mean-square convergence of numerical methods for intrinsically defined SDEs on a Riemannian manifold is studied in Bharath et al. (2025).

Methods convergent in mean-square are needed when one is interested in proximity of trajectories of an SDEs' approximation to trajectories of the exact solution (Milstein and Tretyakov, 2021). However, the task of approximating expectations with respect to SDEs, including that of ergodic limits, is in principle simpler: it was noted in Milstein (1978) that, roughly, weak-sense convergence needs only moments of the SDEs' solution and its approximation to be matched up to some order. Consequently, demands on approximations convergent in weak sense are less stringent than for mean-square schemes, especially in the case of higher-order approximations. In particular on $\mathbb{R}^q$, the Euler scheme for general SDEs has mean-square order $1/2$, while its weak order is 1, and there are constructive weak order schemes of higher orders and no constructive mean-square schemes of higher order than $1/2$ in the general case (Milstein and Tretyakov, 2021).

Thanks to the PDE-based technique for proving weak convergence (Milstein, 1985; Talay, 1986, 1990; Mattingly et al., 2010) (see also Milstein and Tretyakov (2021)), our proofs of weak convergence for Riemannian Langevin algorithms have a unified nature and are easily adaptable to a variety of schemes, for example, involving retraction maps (see Section 3.2). The consequence of this proof technique is that we do not monitor dependence of the coefficients (at h) in upper error bounds on parameters of the problem (such as dimension, curvature, spectral gap, etc). However, using the Talay-Tubaro expansion (Talay and Tubaro, 1990) (see Section 4.4), one can estimate the leading error term of a weak scheme (i.e., the posteriori error), which is valuable in practice.

The central challenge in extending the analysis to non-compact M lies in the fact that, unlike when $M = \mathbb{R}^q$ or a Ricci-flat manifold M (e.g., $\mathbb{T}^q$), it is insufficient to merely control behavior of the drift term $\nabla \phi$, since the non-positive curvature of M also dictates behavior of an intrinsic Brownian motion B^M. The Langevin diffusion can hence fail to be ergodic even when its drift term is well-behaved on a geodesically complete negatively curved M. We discuss conditions that would enable obtaining first-order weak error bounds in Section 5, and also provide numerical illustrations of the same in Section 6.2, but postpone the theoretical analysis to future work.

As mentioned earlier, there are some existing works that provide curvature-dependent bounds for distances between μ_ϕ^h and μ_ϕ under different intrinsic discretizations X_n^h and settings: Grorud and Talay (1996) for chart-dependent Euler scheme (see further discussion at the end of Section 3.1); Cheng et al. (2022) for the geodesic Euler scheme considered

here claim to develop $O(h^{1/2})$ bound in expected squared Riemannian distance between $X(t_n)$ and X_n^h, $t_n = nh$; Gatmiry and Vempala (2022) provide an $O(h)$ bound on the Kullback-Leibler distance for Hessian manifolds, but assume possibility of *exact* sampling of a Brownian increment on M, which currently is available for spheres (Mijatović et al., 2020). Further, the above works do not consider and analyse an estimator of $\mu_\phi(\varphi)$ while we obtain an $O(h)$ bound on the bias of the estimators.

Although not directly related to our work, we note that there are quite a few papers on extrinsic discretization methods for an embedded Langevin diffusion X defined by viewing M as a submanifold of $\mathbb{R}^k, k \geq q$ (e.g., Leliévre et al., 2012; Sharma and Zhang, 2021; Laurent and Vilmart, 2022; Armstrong and King, 2022). A critical limitation of extrinsic methods is that while X is constrained to move on M, the discretized Markov process X_n^h need not, and repeated, computationally expensive, projections on to M are required to ensure that X_n^h does not leave M; an exception is when M is a Lie group, and the group action enables transport along the manifold (e.g., Malham and Wiese, 2008; Davidchack et al., 2009; Mentink et al., 2010; Davidchack et al., 2015). Use of projections results in an additional source of error over the original discretization of X, avoided entirely by intrinsic methods. Intrinsic methods may require additional computation of geodesics when they are unknown in closed-form, but this can be addressed by using computationally cheaper, higher-order retractions which do not affect the theoretical weak error bounds (see Section 3.2). We note that as with the approximation of the projection map in the extrinsic setting, an efficient computation of retraction in our intrinsic setting does not depend on the potential ϕ and noise, and depends only on the manifold.

Projection-based schemes are inappropriate when the target probability density to sample from on a manifold M is defined with respect to a volume form coming from a specific Riemannian metric g that is not the restriction of the standard metric in $\mathbb{R}^k, k \geq q$. Moreover, when M is an *open* submanifold of $\mathbb{R}^k$, in the absence of a level-set representation of M, projection-based schemes are inappropriate. A practically relevant example that highlights the above two issues concerns the manifold $\mathcal{P}_m$ of $m \times m$ real symmetric positive definite matrices of dimension $q = m(m+1)/2$ (see Section 6.2). Firstly, it is an open cone within the same dimensional vector space of symmetric matrices so that $k = m(m+1)/2$; an orthogonal projection would map a symmetric matrix with a single negative eigenvalue to a positive definite matrix with smallest eigenvalue very close to zero, close to the boundary of $\mathcal{P}_m$, which is infinitely far away with respect to the intrinsic geometry of $\mathcal{P}_m$, and the projection-based sampling scheme will fail to sample from the target measure with support in $\mathcal{P}_m$. Secondly, there are numerous choices of intrinsic Riemannian metrics motivated by practical applications (e.g. Pennec et al., 2006, Chapter 3) which lead to specific target probability densities that cannot be sampled with projection-based schemes that views M as a submanifold of $\mathbb{R}^k$ with the induced metric.

Additionally, projection-based schemes are typically less stable (requiring the use of appreciably smaller time steps) compared to geometric integrators, which by construction automatically lie on the manifold; see, for example, Mentink et al. (2010) when $M = \mathbb{S}^2$. For M that is a Lie group or a homogeneous space of one, Lie group integrators preserving the geometry represent an efficient choice. However, for general manifolds M with specific metrics g the intrinsic methods considered in this paper will typically have an advantage when compared with projection-based schemes.

2. Preliminaries

In Section 2.1, we briefly review concepts and fix notation from differential geometry needed for analysis on a Riemannian manifold M; for more details we refer to some well-known sources (e.g., Jost, 2008; Cheeger and Ebin, 1975; Gallot et al., 1990; Bishop and Crittenden, 2011). In Section 2.2, we introduce the Langevin diffusion on M with a given invariant measure, the target measure for sampling.

2.1 Differential geometric concepts

Let M be a q-dimensional topological manifold. By this we mean a Hausdorff, paracompact, locally Euclidean topological space. We further assume that M is connected and without a boundary. A chart of M is a double (U, ψ) where U is an open set of M and $\psi : M \to \mathbb{R}^q$ is a homeomorphism. A collection of these charts that cover M entirely is called an atlas. For any two different charts (U_i, ψ_i) and (U_j, ψ_j) that overlap (for any i, j such that $U_i \cap U_j \neq \emptyset$), the map $\psi_j \circ \psi_i^{-1} : \psi_i(U_i \cap U_j) \to \psi_j(U_i \cap U_j)$ is known as a transition map. Then M is a C^k-differentiable manifold if the transition maps are C^k with C^k inverses. In this paper, we consider the case when $k = \infty$, then M is known as a smooth manifold.

Tangent vectors at a point $x \in M$ may be defined using derivations (linear operators on smooth functions on M compatible with the product rule of differentiation) or as (time) derivatives of an equivalence class of curves passing through x. Denote by $T_x M$ the q-dimensional tangent space of $x \in M$ and by $TM = \cup_x T_x M$ the tangent bundle of M, itself a differentiable manifold of dimension $2q$. A local coordinate system $x = (x^1, \ldots, x^q)^\top$ in a chart induces a basis $\partial_x = \{\partial_{x^1}, \ldots, \partial_{x^q}\}$ at $T_x M$ such that any $v \in T_x M$ may be expressed as $v = v^i \partial_{x^i}$, where the Einstein summation notation, which will be used throughout, is used. The map $X : M \to TM$ is a vector field with $X(x) \in T_x M$ for $x \in M$.

Tangent spaces $T_x M$ and the tangent bundle TM have dual spaces, the cotangent space $T_x^* M$ and cotangent bundle $T^* M$ consisting of linear functionals $\nu : T_x M \to \mathbb{R}$. The coordinate basis $\{\partial_{x^1}, \ldots, \partial_{x^d}\}$ induces a similar basis $\{dx^1, \ldots, dx^q\}$ on $T_x^* M$ defined via the evaluation $dx^j(\partial_{x^i}) = \delta_i^j$, where δ_i^j is the delta function, such that $\nu = \nu_i dx^i$ for every $\nu \in T_x^* M$.

A differentiable manifold M becomes a Riemannian manifold (M, g) when equipped with a Riemannian metric $g : T_x M \times T_x M \to \mathbb{R}$, a family $\{g(x), x \in M\}$ of smoothly varying symmetric positive definite bilinear forms (inner products). In a chart around x, the metric $g(x)$ can be expressed as a symmetric positive definite matrix whose (i, j)-th entry is $g(\partial_{x^i}, \partial_{x^j})$, which we write in index notation as g_{ij}.

Denote by D the unique torsion-free Levi-Civita connection, or covariant derivative, compatible with g. The connection enables definition of a derivative $D_X Y$ of a vector field Y along a vector field X. For a smooth function $f : M \to \mathbb{R}$, denote by $X(f)$ the derivative of f in the direction of X based on the interpretation of tangent vectors as derivations. The covariant derivative is linear in X and follows the product rule in Y so that $D_X(fY) = X(f)Y + f D_X Y$. More conveniently, $D_X Y$ may be written in local coordinates at x using the relation $D_{\partial_{x^i}} \partial_{x^j} = \Gamma_{ij}^k \partial_{x^k}$, where Γ_{ij}^k are the Christoffel symbols.

The length of a curve $c : [0, 1] \to M$ connecting $c(0) = x$ and $c(1) = y$ is given by

$$L(c) = \int_0^1 \sqrt{g(\dot{c}(t), \dot{c}(t))} dt,$$

7

where $\dot{c} \in T_{c(t)}M$ is the tangent vector to the curve at $c(t)$. A curve that minimises the length functional is called a *geodesic*, and this may not be unique. A geodesic γ satisfies the equation $D_{\dot{\gamma}}\dot{\gamma} = 0$, and in local coordinates may be expressed as

$$\frac{\mathrm{d}^2\gamma^k}{\mathrm{d}t^2} + \Gamma^k_{ij}\frac{\mathrm{d}\gamma^i}{\mathrm{d}t}\frac{\mathrm{d}\gamma^j}{\mathrm{d}t} = 0. \tag{2.1}$$

Geodesics enjoy a rescaling property: if $t \mapsto \gamma(t)$ is a geodesic with $\gamma(0) = x$ and $\dot{\gamma}(0) = v$, then $t \mapsto \gamma(\alpha t)$ is a geodesic with $\dot{\gamma}(0) = \alpha v$ for every $(x, v) \in TM$ and $\alpha, t \in \mathbb{R}$. Aided by the rescaling property of geodesic, the Riemannian distance $(x, y) \mapsto \rho(x, y)$ between two points $x, y \in M$ is defined to be the length of the geodesic γ with $\gamma(0) = x$ and $\gamma(1) = y$.

A manifold M is said to be geodesically complete if every geodesic $\gamma : (-a, a) \to M$ can be extended to a geodesic with domain $\mathbb{R}$, and by the Hopf-Rinow theorem, the space (M, ρ) is complete as a metric space, and the two notions of completeness coincide; this ensures that for such manifolds there exists at least one minimizing geodesic between any two points whose length is the distance between the points. Moreover, (M, ρ) is said to be compact if it is a compact metric space, and every compact Riemannian manifold without boundary is complete.

Theory of differential equations guarantees that for every $(x, v) \in TM$, there exists a unique geodesic γ starting at x with $\gamma(0) = x$ with $\dot{\gamma}(0) = v$. This results in a surjective map on T_xM for complete M that maps a vector $v \in T_xM$ to a point on M along geodesics: straight lines through the origin in T_xM map to geodesics through x on M.

Definition 1 *Let $\gamma : [0, 1] \to M$ be a geodesic with $\gamma(0) = x$ and $\dot{\gamma}(0) = v$. Then*

$$\exp_x : T_xM \to M, \quad v \mapsto \exp_x(v) := \gamma(1)$$

is known as the exponential map at x.

Although surjectivity ensures that $\exp_x$ is defined on all of T_xM for every $x \in M$, it is injective only in a neighbourhood N_x around the origin in T_xM. Every geodesic $\gamma(tv) = \exp_x(tv)$ starting from $t = 0$ is either minimizing on all $t \in \mathbb{R}$ or is minimizing until $s < \infty$; then $\gamma(sv)$ is called a *cut point* of x. The set of all cut points of all geodesics starting from x is the cut locus cut_x of x.

The neighbourhood N_x is star-shaped and connected, and $\exp_x(N_x) = M - \mathrm{cut}_x$. Relatedly, the radius of injectivity, $\mathrm{inj}(x)$, of a point $x \in M$ is the maximal radius of balls around the origin in T_xM on which the exponential map is injective; or, in other words, it is the radius r of the largest geodesic ball centered at x such that each geodesic connecting to x within the ball is minimizing. The injectivity radius inj_M of the manifold M is the smallest of such radii across M.

It is necessary to introduce a handful of tensors in preparation for the analysis of the algorithm. First, the exterior derivative d is a differential operator that maps C^1 functions to the cotangent bundle T^*M, and in local coordinates assumes the form $\mathrm{d}f(x) = \frac{\partial f(x)}{\partial x^i}\mathrm{d}x^i$. The quantity $\mathrm{d}f$ is known as a 1-form. More generally, an r-form, for $r \in \{1, \ldots, q\}$, can be considered on $\wedge^r T^*_x M$ with anti-symmetric wedge product $\wedge$ spanned by $\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^r$. For brevity, the wedge symbol is omitted when discussing differential forms. Then, the exterior derivative d maps r-forms to $(r+1)$-forms and $\mathrm{d}^2 = 0$. For the purposes of integration,

we introduce the volume form, a q-form, $\mathrm{dvol}_g = \sqrt{|g|}\mathrm{d}x^1\ldots\mathrm{d}x^q$, where $|g|$ denotes the determinant of the Riemannian metric g in local coordinates around x.

For $v \in T_xM$, the relation

$$\mathrm{d}f(x)(v) = g(\nabla f(x), v)$$

defines the Riemannian gradient operator ∇ acting on smooth functions $f : M \to \mathbb{R}$ as the dual of $\mathrm{d}f$. In local coordinates, the gradient is written as $\nabla f = g^{ij}\frac{\partial f}{\partial x^i}\frac{\partial}{\partial x^j}$ where g^{ij} are the elements of the inverse metric; it is often convenient to omit dependence on x and use $\mathrm{d}f(v), v \in T_xM$.

The L^2-adjoint of the gradient is the negative of the divergence operator. For a tangent vector $v \in T_xM$ the local coordinate representation of divergence is $\mathrm{div}(v) = \frac{1}{\sqrt{|g|}}\frac{\partial}{\partial x^i}(\sqrt{|g|}v^i)$.

The Hessian of a C^2 function $f : M \to \mathbb{R}$ is

$$\mathrm{Hess}^f = D^2 f = D\mathrm{d}f.$$

Then, the Laplace-Beltrami operator, as an analogue of the Laplacian in $\mathbb{R}^q$, is defined as the trace (with respect to the metric tensor) of the Hessian

$$\Delta_M f = \mathrm{TrHess}^f = g^{ij}\mathrm{Hess}^f_{ij}.$$

For higher derivatives of functions, $D^k : C^\infty(M) \to (T^*M)^{\otimes k}$, i.e. D^k maps smooth functions to the k times cotangent space. We can contract $D^k f$ by interpreting it as the mapping $D^k f : (TM)^{\otimes k} \to \mathbb{R}$. For the contracted tensor $S \in (T^*M)^{\otimes k}$, $S(v_1,\ldots,v_k)$ for arbitrary $v \in TM$, the covariant derivative acts multilinearly on the $(0,k)$-tensor S and also on each of the vector fields. More precisely, $D_v(S(v_1,\ldots,v_k)) = (D_vS)(v_1,\ldots,v_k) + S(D_vv_1, v_2,\ldots,v_k) + \cdots + S(v_1,\ldots,v_{k-1}, D_vv_k)$ (e.g., Jost, 2008, pp. 202). The norm g can be extended to tensors, for a tensor $S \in T_x^{p,q}M$ the norm of S is

$$|S|^2 = g_{i_1k_1}\cdots g_{i_pk_p}g^{j_1l_1}\cdots g^{j_ql_q}S^{i_1\cdots i_p}_{j_1\cdots j_q}S^{k_1\cdots k_p}_{l_1\cdots l_p}.$$

The parallel transport $\Pi_{\gamma_{x,y}} : T_xM \to T_yM$ along a curve $\gamma_{x,y} : [0,1] \to M$ connecting x and y is a linear isometry determined by a unique vector field X satisfying $D_{\dot\gamma(t)}X = 0$ for all t, such that $g(\Pi_{\gamma_{x,y}}u, \Pi_{\gamma_{x,y}}v) = g(u,v)$ for all $u, v \in T_xM$. More generally, the parallel transport $\Pi_{\gamma_{x,y}}$ of a tensor $S \in T_x^{p,q}M$ along an arbitrary geodesic (or even curve) γ from the point $x = \gamma(0)$ to $y = \gamma(1)$ is such that it preserves the inner product between S and $\dot\gamma$, $g(S, \dot\gamma(0)) = g(\Pi_{\gamma_{x,y}}S, \dot\gamma(1))$, and also its norm, $g(S,S) = |S|^2 = |\Pi_{\gamma_{x,y}}S|^2$. It is similarly defined by solving the ODE $D_{\dot\gamma}\Pi S = 0$.

Define the operator norm for a $(0,k)$-tensor $\mathfrak{T}$ on M (e.g. Le et al., 2024)

$$|\mathfrak{T}|_{\mathrm{op}} := \sup_{V_1,\ldots,V_k\in T_xM,\,|V_i|\neq 0} \frac{|\mathfrak{T}(V_1,\ldots,V_k)|}{\prod_{i=1}^k |V_i|} \tag{2.2}$$

for each $x \in M$. In addition to spaces of continuously differentiable functions $C^l(M)$, we will need Hölder spaces $C^{l,\epsilon}(M)$, $0 < \epsilon < 1$, containing real-valued functions $u(x)$, $x \in M$, with corresponding number of Hölder continuous derivatives and with the norm

$$\|u\|_{C^{l,\epsilon}(M)} := \sum_{k=0}^l \sup_{x\in M}|D^ku(x)|_{\mathrm{op}} + \sup_{\gamma_{x,y},x\neq y\in M} \frac{|D^lu(x) - \Pi_{\gamma_{x,y}}(D^lu(y))|_{\mathrm{op}}}{(\rho(x,y))^\epsilon},$$

where $\gamma_{x,y}$ denotes any possible minimal geodesic from y to x and $\Pi_{\gamma_{x,y}}$ denotes the parallel transport from $T_y M$ to $T_x M$ along $\gamma_{x,y}$. Analogously, $C^{l/2,l,\epsilon}([0,T] \times M)$ is a space of sufficiently smooth functions $u(t,x)$ with the norm for even l:

$$\|u\|_{C^{l/2,l,\epsilon}([0,T]\times M)} := \sum_{k=0}^{l/2} \sum_{n=0}^{l-2k} \sup_{t\in[0,T]} \sup_{x\in M} \left| D^n \frac{\partial^k}{\partial t^k} u(t,x) \right|_{\text{op}}$$

$$+ \sum_{k=0}^{l/2} \sup_{t\in[0,T]} \sup_{\gamma_{x,y}, x\neq y\in M} \frac{|D^{l-2k}\frac{\partial^k}{\partial t^k}u(t,x) - \Pi_{\gamma_{x,y}}(D^{l-2k}\frac{\partial^k}{\partial t^k}u(t,y))|_{\text{op}}}{(\rho(x,y))^\epsilon}.$$

For vector fields X,Y,Z, the Riemannian curvature tensor at a point $x \in M$ $R: T_x M \times T_x M \times T_x M \to T_x M$ of M is defined as

$$R(X,Y)Z := (D_X D_Y - D_Y D_X - D_{[X,Y]})Z,$$

where the commutativity bracket (or Lie bracket) is defined via its application to a smooth function $f: M \to \mathbb{R}$ by $[X,Y]f = X(Y(f))-Y(X(f))$. Roughly, it quantifies the difference between the vectors obtained by transporting a vector $Z(x)$ first along $X(x)$ then $Y(x)$, in contrast to moving first along $Y(x)$ and then $X(x)$. The sectional curvature at x is the quadratic form (Hsu, 2002, pp. 89)

$$K(X,Y) := g(R(X,Y)Y,X), \quad X,Y \in T_x M,$$

and for an orthonormal basis $\{X_i(x)\}$ on $T_x M$, the Ricci curvature at x is a contraction of the sectional curvature:

$$\text{Ric}(X,X) := \sum_{i=1}^q g(R(X,X_i)X_i,X).$$

2.2 Langevin diffusion on a Riemannian manifold

Let (M,g) be a connected compact Riemannian manifold of dimension $q \geq 1$ without boundary; this will be the main setting assumed throughout the paper except in Sections 5 and 6.2. Consider the target probability measure μ_ϕ on M

$$\mathrm{d}\mu_\phi = p\,\mathrm{dvol}_g = \frac{1}{C_\phi}e^{-\phi}\mathrm{dvol}_g, \tag{2.3}$$

absolutely continuous with density p with respect to the volume measure dvol_g, where $C_\phi = \int_M e^{-\phi}\mathrm{dvol}_g < \infty$. The following conditions on smoothness of the metric g and the density p are imposed through that of ϕ.

Assumption 1 *Assume that $x \mapsto g^{ij}(x) \in C^{3,\epsilon}(M)$.*

Assumption 2 *Assume that $\phi \in C^{3,\epsilon}(M)$.*

No further restrictions are placed on ϕ (e.g., convexity). The level of smoothness imposed in these assumptions is required primarily for proofs of Theorems 1-3, but may be weakened for the content of this section.

Since we aim to sample from a given probability measure μ_ϕ, we must find an (second order) elliptic linear operator $\mathcal{A}^*$ acting on probability densities p (with respect to dvol_g) so that the density p from (2.3) satisfies the stationary Fokker-Planck equation (Ikeda and Watanabe, 2014, Chapter 5, Proposition 4.5):

$$\mathcal{A}^* p = 0. \tag{2.4}$$

For example, one can verify that the above is satisfied by the operator

$$\mathcal{A}^* p = \frac{1}{2}\Delta_M p + \frac{1}{2}\mathrm{div}(p\nabla\phi), \tag{2.5}$$

which is adjoint to the elliptic operator (Ikeda and Watanabe, 2014, Chapter 5, Proposition 4.4)

$$\mathcal{A} = \frac{1}{2}\Delta_M - \frac{1}{2}g(\nabla\phi, \nabla) \tag{2.6}$$

acting on $C^2(M)$-functions, where Δ_M is the Laplace-Beltrami operator.

The Markov process with infinitesimal generator $\mathcal{A}$ is governed by the following SDE (e.g., Elworthy, 1982; Ikeda and Watanabe, 2014)

$$\mathrm{d}X(t) = -\frac{1}{2}\nabla\phi(X(t))\mathrm{d}t + \mathrm{d}B^M(t), \quad 0 \leq t \leq T;\ X(0) = x, \tag{2.7}$$

where B^M is a Brownian motion on the manifold M defined using the Eells-Elworthy-Malliavin frame bundle construction (e.g., Hsu, 2002, p.87).

The Brownian motion B^M can be intuited as follows: given a filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathbb{P}), t \geq 0$, and an $\{\mathcal{F}_t\}_{t>0}$-adapted $\mathbb{R}^q$-valued standard Brownian motion B, the construction uses the dynamics of B on $\mathbb{R}^q$ to induce one on M by first, mapping the dynamics of B onto the orthonormal frame bundle using the notion of horizontal vector fields, and then projecting down from the frame bundle onto the manifold. Thus, in local coordinates in a chart the SDE (2.7) assumes the Itô form

$$\mathrm{d}X^i(t) = -\frac{1}{2}g^{ij}(X(t))\frac{\partial\phi}{\partial x^j}(X(t))\mathrm{d}t - \frac{1}{2}g^{kj}(X(t))\Gamma_{kj}^i(X(t))\mathrm{d}t + (g^{1/2}(X(t)))^{ij}\mathrm{d}B_j(t), \tag{2.8}$$

$$X(0) = x,$$

where B_j are $\mathbb{R}$-valued components of B. Under Assumptions 1 and 2, (2.8) is well-defined. The frame $g^{-1/2}(x) : \mathbb{R}^q \to T_x M$ maps Brownian dynamics from $\mathbb{R}^q$ to $T_x M$, and this leads to the diffusion coefficient $g^{-1/2}$ in (2.8).

On a Riemannian manifold, not necessarily compact, it is known (Bakry, 1986) that the diffusion X does not explode, i.e., it is defined for any $t \geq 0$, if

$$\mathrm{Ric}(v, v) + \mathrm{Hess}^\phi(v, v) \geq -\kappa g(v, v) \ \ \text{for all}\ (x, v) \in TM \tag{2.9}$$

for some $\kappa > 0$. The condition is satisfied for any compact M, since the Hessian of ϕ is bounded under Assumption 2; moreover, the Ricci curvature is bounded from below since the sectional curvature at any point in M is bounded from below (Bishop and Crittenden,

2011, pp. 167). We will revisit this condition in Section 5 on non-compact manifolds, where we also discuss how it relates to the Euclidean case.

A key requirement for sampling and estimation using $X(t)$ is its ergodicity. Recall that a diffusion $X(t)$ is ergodic if there exists a unique invariant measure μ of $X(t)$, and independently of the initial condition $x \in M$, the limit

$$\lim_{t \to \infty} \mathbb{E}\varphi(X_x(t)) = \int_M \varphi(x)\, \mathrm{d}\mu_\phi(x) =: \mu_\phi(\varphi) \tag{2.10}$$

exists for any bounded and measurable function $\varphi : M \to \mathbb{R}$. Furthermore, the process $X(t)$ is exponentially ergodic if for any $x \in M$ and any bounded function φ we have the following strengthening of (2.10):

$$|\mathbb{E}\varphi(X_x(t)) - \mu_\phi(\varphi)| \le Ce^{-\lambda t}, \quad t \ge 0, \tag{2.11}$$

where $C > 0$ and $\lambda > 0$ are some constants. Since the process $X(t)$ is a non-degenerate diffusion on a compact smooth manifold M, its distribution converges at an exponential rate to the invariant probability measure μ_ϕ (Freidlin, 1985, p. 97). Consequently, we can exploit the SDE (2.7) (or its local version (2.8)) for computing ergodic limits $\mu_\phi(\varphi)$.

Consider two examples of Langevin diffusions on compact M.

Example 1 *Let $M = \mathbb{S}^q$ with canonical round metric $g = \mathrm{d}r^2 + \sin^2 r\, \mathrm{d}\theta^2$ with intrinsic coordinate $x = (r, \theta)$ for $r \in [0, \pi)$ and $\theta \in \mathbb{S}^{q-1}$. The Fisher-Watson distribution has the probability density function $p(x) \propto \exp(\lambda \cos^2 r)$ with respect to the volume form for $\lambda > 0$, $r \in [0, \pi)$ (Watson, 1965), and thus $\phi(x) = -\lambda \cos^2 r$, which clearly is in $C^\infty(M) \subset C^{3,\epsilon}(M)$. At least, two charts are needed for the local representation of X since the cut locus of every point is its antipode, and non-empty.*

Example 2 *Let $M = \mathrm{SO}(m)$ be the space of real $m \times m$ orthogonal matrices with determinant 1, a manifold of dimension $q = m(m-1)/2$. Consider the bi-invariant metric $g(E_1, E_2) = -\frac{1}{2}\mathrm{Tr}(E_1 E_2)$ for $x \in \mathrm{SO}(m)$, where $E_1, E_2 \in T_x\mathrm{SO}(m) \cong \{x\} \times \mathfrak{so}(m)$ and where $\mathfrak{so}(m)$ is the Lie algebra of $\mathrm{SO}(m)$ containing skew-symmetric matrices. The matrix generalisation of the von-Mises Fisher distribution, known as the matrix von-Mises distribution (Downs, 1972; Jupp and Mardia, 1979), has the density function $p(x) \propto \exp(c\mathrm{Tr}(x_0 x))$ for $c > 0$ and $x, x_0 \in \mathrm{SO}(m)$. With $\phi = -c\mathrm{Tr}(x_0 x)$ it can be verified that for $x \in \mathrm{SO}(m)$ and $e \in \mathfrak{so}(m)$, $\mathrm{Hess}^\phi(xe, xe) = -c\mathrm{Tr}(xe^2)$, which shows that $\phi \in C^2(M)$ (Lewis, 2023, Section 4.3).*

2.3 Partial differential equations on M related to the Langevin diffusion

Central to our analysis of estimators of a linear functional $\mu_\phi(\varphi)$ of the invariant measure μ_ϕ are two PDEs that link the generator $\mathcal{A}$ to $\mu_\phi(\varphi)$.

2.3.1 BACKWARD KOLMOGOROV PDE

The backward Kolmogorov PDE on manifold M associated with the SDE (2.7) (e.g., Ikeda and Watanabe, 2014) is given by:

$$\frac{\partial u}{\partial t}(t, x) + \mathcal{A}u(t, x) = 0, \quad (t, x) \in [0, T] \times M;$$

$$u(T, x) = \varphi(x), \quad x \in M. \tag{2.12}$$

Under Assumptions 1 and 2 and $\varphi \in C^{4,\epsilon}(M)$, there is a unique solution of the parabolic problem (2.12) which belongs to $C^{2,4,\epsilon}([0, T] \times M)$ (Aubin, 1998; Ladyzhenskaya et al., 1968; Lunardi, 1995; Ikeda and Watanabe, 2014) and

$$\|u\|_{C^{2,4,\epsilon}([0,T] \times M)} \le C\|\varphi\|_{C^{4,\epsilon}(M)}, \tag{2.13}$$

where $C > 0$ is a constant independent of φ. The solution of (2.12) also has the following property (Grorud and Talay, 1996, Lemma 6.4), which links geometric ergodicity of the Markov process $X(t)$ solving (2.7) to the decay of the norm of derivatives of its semigroup: there exist constants $\lambda > 0$ and $C > 0$ that do not depend on T and choice of φ such that

$$\sum_{k=0}^{2} \sum_{\substack{n=0 \\ k+n \ne 0}}^{4-2k} \left\| \frac{\partial^k}{\partial t^k} u(t, \cdot) \right\|_{C^{n,\epsilon}(M)} \le C\|\varphi\|_{C^{4,\epsilon}(M)} e^{-\lambda(T-t)}. \tag{2.14}$$

For every $(t_0, x_0) \in [0, T] \times M$, the Feynman-Kac formula (Ikeda and Watanabe, 2014; Freidlin, 1985) provides a probabilistic representation of the solution of (2.12) by linking it to the law of the Langevin diffusion $X(t)$ via

$$u(t_0, x_0) = \mathbb{E}(\varphi(X(T))), \tag{2.15}$$

where X is from (2.7) with initial condition $X(t_0) = x_0$. The representation enables construction of the ensemble-averaging estimator of $\mu_\phi(\varphi)$ using multiple independent trajectories of the discretized X_n^h.

2.3.2 POISSON PDE

The Poisson PDE on M associated with the SDE (2.7) has the form

$$\mathcal{A}u(x) = \varphi(x) - \mu_\phi(\varphi), \quad x \in M. \tag{2.16}$$

It is clear that $\int_M (\varphi(x) - \mu_\phi(\varphi)) \mathrm{d}\mu_\phi(x) = 0$, i.e., the centering condition is satisfied (e.g., Miranda, 1969; Freidlin, 1985). Under Assumptions 1 and 2 and $\varphi \in C^{2,\epsilon}(M)$, there is a unique (upto an additive constant) solution of the elliptic problem (2.16) which belongs to $C^{4,\epsilon}(M)$ (Aubin, 1998; Nicolaescu, 2007; Miranda, 1969) and

$$\|u\|_{C^{4,\epsilon}(M)} \le C\|\varphi\|_{C^{2,\epsilon}(M)}, \tag{2.17}$$

where $C > 0$ is a constant independent of a choice of φ.

Using Itô's formula, one can derive the probabilistic representation of solution of (2.16) as (e.g., Freidlin, 1985; Le et al., 2024)

$$u(x) = -\mathbb{E} \int_0^\infty \big(\varphi(X(t)) - \mu_\phi(\varphi)\big) \mathrm{d}t + \mu_\phi(u).$$

The representation motivates an estimator of $\mu_\phi(\varphi)$ based on a long single trajectory of X_n^h, which can then be analysed by the mean ergodic theorem.

3. Langevin Sampling Algorithms

Section 3.1 introduces the intrinsic geodesic Euler discretization of the intrinsic Langevin SDE (2.7) to sample from a given distribution μ_ϕ on a Riemannian manifold. The resulting algorithm can be used on non-compact M as well, and numerical experiments with the non-compact manifold of symmetric positive definite matrices in Sections 6.2.1 and 6.2.2 demonstrate this. Section 3.2 generalises the Riemannian Langevin algorithm to one based on retractions, of which the exponential map is a special case. Section 3.3 discusses possible variants and extensions of the Riemannian Langevin algorithm.

3.1 Riemannian Langevin algorithm

For $T > 0$ and $t_0 = 0$, consider a uniform partition $t_0 < \cdots < t_N = T$ of the interval $[0, T]$ with time step $h := T/N$, i.e. $t_{n+1} - t_n = h$, $n = 0, \ldots, N - 1$. In $\mathbb{R}^q$, the Euler method to discretize (1.1) assumes the form (see e.g. Milstein and Tretyakov (2021))

$$X_{n+1}^h = X_n^h - \frac{h}{2}\nabla\phi(X_n^h) + \sqrt{h}\xi_{n+1},$$

where ξ_n, $n = 1, \ldots, N$, are independent standard Gaussians on $\mathbb{R}^q$ such that $\sqrt{h}\xi_n$ represents a Brownian increment in a small time step of size h. Evidently, the dynamics are then along the straight line from X_n^h with respect to the direction

$$v_{n+1} := -\frac{h}{2}\nabla\phi(X_n^h) + \sqrt{h}\xi_{n+1}.$$

Its natural generalisation to the curved M is via a Markov process whose next state moves along a geodesic with a direction chosen in the tangent space of the current state. Upon isometrically identifying the tangent space with $\mathbb{R}^q$, the following two-step procedure ensues:

(i) In the tangent space $T_{X_n^h}M$, select a random direction v_{n+1} via an $\mathbb{R}^q$-valued random vector ξ_{n+1};

(ii) follow the geodesic $[0, 1] \ni t \mapsto \gamma(t) \in M$ with $\gamma(0) = X_n^h, \dot{\gamma}(0) = v_{n+1}$ to $t = 1$, and set $X_{n+1}^h := \gamma(1)$.

Note that geodesics can be parameterized using the exponential map $TM \ni (x, v) \mapsto \exp_x(v) \in M$ defined in Section 2.1 such that $\gamma(1) = \exp_{X_n^h}(v_{n+1})$. Surjectivity of the exponential map on complete manifolds M ensures that the two-step procedure is well-defined without requiring any restriction on the support of distribution of the random vector ξ_{n+1}. It is possible to replace the geodesic with curves arising from retractions of suitable order without affecting the order of convergence of the algorithm (see Section 3.2).

Choice of the distribution of $\{\xi_n\}$ may be of practical consequence (Section 6.2). Nevertheless, when studying weak convergence, it suffices to consider a distribution that matches k moments of a Gaussian, where k is related to the order of weak convergence (Milstein and Tretyakov, 2021) and in the case of first order weak methods like the Euler scheme it is sufficient to match the first three moments. Thus, in contrast to the algorithms currently available in the literature (e.g., Cheng et al., 2022) which use Gaussian distributed

ξ_n in $T_{X_n^h}$ with respect to the inner product $g(X_n^h)$, we only require that the components of $\xi_n^i, i = 1, \ldots, q$ satisfy

$$\mathbb{E}[\xi_n^i] = 0, \quad \mathbb{E}[(\xi_n^i)^2] = 1, \quad \mathbb{E}[(\xi_n^i)^3] = 0, \quad \mathbb{E}[(\xi_n^i)^4] < \infty \tag{3.1}$$

for every $n = 0, \ldots, N - 1$. It suffices, for example, to use a discrete distribution so that for each $n = 0, \ldots, N - 1$,

$$P(\xi_n^i = \pm 1) = \frac{1}{2}, \quad i = 1, \ldots, q. \tag{3.2}$$

The Markov chain $\{X_n^h, n = 0, 1, \ldots, N\}$ with $X_0^h = x$, defined as

$$X_{n+1}^h = \exp_{X_n^h}\left(-\frac{h}{2}\nabla\phi(X_n^h) + h^{1/2}g^{-1/2}(X_n^h)\xi_{n+1} \right), \tag{3.3}$$

represents the discretization of (2.7). The intrinsic method of discretization without using an embedding into $\mathbb{R}^k$, $k \geq q$, ensures that it suffices to use a q-dimensional random vector ξ_n, as opposed to a higher-dimensional one used in extrinsic approaches (e.g. Laurent and Vilmart, 2022; Sharma and Zhang, 2021; Armstrong and King, 2022).

Importantly, use of the exponential map, or more generally, retractions (see Section 3.2), ensures that X_n^h always lies on the manifold M. As discussed in the Introduction, this feature of the algorithms discussed here and in the next subsection is important for their long time stability (in terms of being able to use larger time steps h and hence resulting in overall faster simulations), which is crucial for efficient sampling.

The proposed Riemannian Langevin algorithm may be fruitfully compared to the one proposed by Grorud and Talay (1996). In their paper an Euler scheme was defined in local coordinates of a chart, and the random variables were constrained to have compact support within the chart to ensure that X_n^h and X_{n+1}^h lie in the same chart for sufficiently small h, to facilitate proof of weak convergence of the scheme. In contrast, our algorithm, and the subsequent proof of weak convergence, requires no such constraints (beyond the natural moment matching condition (3.1)) since the Markov chain arising from our algorithm moves directly along geodesics, and there is no need to verify that X_n^h and X_{n+1}^h belong to the same chart. In Section 6.1, we experimentally compare the use of Gaussian and discrete random variables within (3.3).

3.2 Retractions

Closed form expressions for the exponential maps are available for very few manifolds (M, g). Practical utility of the Riemannian Langevin algorithm may be enhanced by using more general retraction maps $F_x : T_xM \to M$ in place of the exponential map. Then the update rule for the Markov chain becomes

$$X_{n+1}^h = F_{X_n^h}\left(-\frac{h}{2}\nabla\phi(X_n^h) + h^{1/2}g^{-1/2}(X_n^h)\xi_{n+1} \right). \tag{3.4}$$

A *retraction* is a smooth map from the tangent bundle $F : TM \to M$, whose restriction F_x to T_xM satisfies: (i) $F_x(0) = x$; (ii) $DF_x(0) = \mathrm{id}_{T_xM}$, the identity mapping on T_xM. In the

definition, 0 is the origin of the concerned tangent space. We associate with the retraction $F_{X_n^h}$ a curve c which satisfies the initial conditions

$$c(0) = X_n^h, \tag{3.5}$$

$$\dot{c}(0) = V := -\frac{\sqrt{h}}{2}\nabla\phi(x) + g^{-1/2}(x)\xi. \tag{3.6}$$

Then by (3.4), $c(\sqrt{h}) = X_{n+1}^h$.

There are several ways to construct retractions. One way of obtaining a computationally inexpensive retraction is via numerical approximation of the geodesic equation. The geodesic equation (2.1) takes the following form in local coordinates:

$$\begin{bmatrix} \dot{\gamma} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} v \\ -\Gamma_{ij}v^i v^j \end{bmatrix} \tag{3.7}$$

with initial condition

$$\begin{bmatrix} \gamma(0) \\ v(0) \end{bmatrix} = \begin{bmatrix} x \\ V \end{bmatrix}, \tag{3.8}$$

where

$$V = -\frac{\sqrt{h}}{2}\nabla\phi(x) + g^{-1/2}(x)\xi.$$

To obtain the first weak order of convergence of the Riemannian Langevin algorithm (3.4) as in the case of the exponential map (3.3), it is necessary to impose conditions on the retraction. This relates to the leading h^2 terms in the local error expansion of Lemma B.2. Consequently, we make the following assumption.

Assumption 3 *For a curve $c : [0, \sqrt{h}] \to M$ with initial conditions (3.5)-(3.6) corresponding to the retraction F_x, we require*

$$|\mathbb{E}[D_{\dot{c}(s)}\dot{c}(s)|_{s=0}]| \leq Ch$$

and

$$|\mathbb{E}[D_{\dot{c}(s)}D_{\dot{c}(s)}\dot{c}(s)|_{s=0}]| \leq Ch^{1/2},$$

where $C > 0$ is independent of h.

To illustrate our retraction based approach, we solve the geodesic equation (3.7) by using a single step of size $\sqrt{h}$ of the standard $(1/6, 1/3, 1/3, 1/6)$ 4th-order Runge-Kutta method (RK4) (see e.g. Hairer et al. (1993)). This is employed in Section 6.1 on the 2-sphere $\mathbb{S}^2$, where the exponential map is not readily available in analytic form in the intrinsic coordinates (r, θ) for the round metric. Since this approximation occurs within a specific, localized coordinate system, the resulting solution belongs to the manifold M, and hence a projection onto M is not required, i.e., there is no additional loss in accuracy.

We now verify that the curve obtained by applying RK4 method to (3.7) with time step $s \in (0, \sqrt{h}]$ satisfies Assumption 3. We denote

$$p(s) = \begin{bmatrix} c(s) \\ \tilde{c}(s) \end{bmatrix} := \begin{bmatrix} \gamma(s) \\ v(s) \end{bmatrix} + e(s) = \begin{bmatrix} \gamma(s) \\ v(s) \end{bmatrix} + \begin{bmatrix} e_1(s) \\ e_2(s) \end{bmatrix}, \tag{3.9}$$

16

where $|e_i(s)| \leq Cs^5$, $i = 1, 2$. Note that, since the curve $p(s)$ in TM corresponds to the one step of RK4 method with step s, we can write the error $e_1(s)$ as the following expansion (Hairer et al., 1993) in terms of s:

$$e_1(s) = s^5 a_1 + s^6 a_2 + s^7 a_3 + O(s^8), \tag{3.10}$$

where $a_i \in \mathbb{R}^q$, $i = 1, \ldots, 3$, are independent of s.

Writing $D_{\dot{c}(s)}\dot{c}(s)$ in local coordinates we obtain

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2}c^k + \Gamma_{ij}^k \frac{\mathrm{d}c^i}{\mathrm{d}s}\frac{\mathrm{d}c^j}{\mathrm{d}s} = \frac{\mathrm{d}^2}{\mathrm{d}s^2}\gamma^k + \Gamma_{ij}^k \frac{\mathrm{d}\gamma^i}{\mathrm{d}s}\frac{\mathrm{d}\gamma^j}{\mathrm{d}s} + \frac{\mathrm{d}^2}{\mathrm{d}s^2}e_1^k + \Gamma_{ij}^k \Big(\frac{\mathrm{d}e_1^i}{\mathrm{d}s}\frac{\mathrm{d}e_1^j}{\mathrm{d}s} + \frac{\mathrm{d}\gamma^i}{\mathrm{d}s}\frac{\mathrm{d}e_1^j}{\mathrm{d}s} + \frac{\mathrm{d}e_1^i}{\mathrm{d}s}\frac{\mathrm{d}\gamma^j}{\mathrm{d}s}\Big)$$

$$= \frac{\mathrm{d}^2}{\mathrm{d}s^2}e_1^k + \Gamma_{ij}^k \Big(\frac{\mathrm{d}e_1^i}{\mathrm{d}s}\frac{\mathrm{d}e_1^j}{\mathrm{d}s} + \frac{\mathrm{d}\gamma^i}{\mathrm{d}s}\frac{\mathrm{d}e_1^j}{\mathrm{d}s} + \frac{\mathrm{d}e_1^i}{\mathrm{d}s}\frac{\mathrm{d}\gamma^j}{\mathrm{d}s}\Big),$$

where $k = 1, \ldots, q$. It is then clear that $D_{\dot{c}(s)}\dot{c}(s)|_{s=0} = 0$ a.s.. Let (c, ϑ) be a curve on TM such that $\vartheta(s) \in T_{c(s)}M$ and $k-$th component of $\vartheta(s)$ is given by

$$\vartheta^k(s) := \frac{\mathrm{d}^2}{\mathrm{d}s^2}e_1^k + \Gamma_{ij}^k \Big(\frac{\mathrm{d}e_1^i}{\mathrm{d}s}\frac{\mathrm{d}e_1^j}{\mathrm{d}s} + \frac{\mathrm{d}\gamma^i}{\mathrm{d}s}\frac{\mathrm{d}e_1^j}{\mathrm{d}s} + \frac{\mathrm{d}e_1^i}{\mathrm{d}s}\frac{\mathrm{d}\gamma^j}{\mathrm{d}s}\Big), \tag{3.11}$$

and we are interested in $D_{\dot{c}(s)}D_{\dot{c}(s)}\dot{c}(s) = D_{\dot{c}(s)}\vartheta(s)$. We can write $D_{\dot{c}(s)}\vartheta(s)$ in local coordinates as

$$\frac{\mathrm{d}}{\mathrm{d}s}\vartheta^k + \Gamma_{ij}^k \vartheta^i \frac{\mathrm{d}}{\mathrm{d}s}c^j(s),$$

which is of order $O(s^2)$ since all the terms have a derivative of e_1 and the highest derivative is of 3rd order (see (3.10)). Consequently, $D_{\dot{c}(s)}D_{\dot{c}(s)}\dot{c}(s)|_{s=0} = 0$ a.s.. Thus, the approximation of the geodesic equation (2.1) by the one step of RK4 with time step $\sqrt{h}$ satisfies Assumption 3.

For other applications of retractions, see Schwarz et al. (2023), where retraction-based geodesic random walks on compact Riemannian manifolds were studied, and Absil et al. (2008), where retractions are used in optimization algorithms on manifolds.

3.3 Other Langevin-based sampling algorithms

It is always beneficial to have a range of algorithms for solving problems numerically. Here we consider two further variants of the Riemannian Langevin algorithm, each of which again ensures that X_n^h always stays on M. Similarly to the algorithms from the previous two subsections, these algorithms are likewise of weak order 1 (see Remark 1).

- Consider different random variables than ξ_n.

  An example of which is the generalisation of the *walk-on-sphere* (Milstein and Tretyakov, 2021, Section 7.1.1) on $\mathbb{R}^q$ to Riemannian manifolds given by

  $$X_{n+1}^h = \exp_{X_n^h}\Big(-\frac{h}{2}\nabla\phi(X_n^h) + \sqrt{qh}g^{-1/2}(X_n^h)\eta_{n+1}\Big), \tag{3.12}$$

  where $\{\eta_n\}$ are independent uniformly distributed on the unit sphere in $\mathbb{R}^q$, and $\sqrt{q}$ is used to take into account that $\mathbb{E}(\eta_n^i \eta_m^j) = \delta_{ij}/q$ so that $\sqrt{q}\eta_n^i$ satisfy the conditions on moments (3.1).

- Consider *splitting methods* that divide each iteration of the sampling algorithm into first approximating the drift $\nabla\phi$ followed by approximating a Brownian increment.

An example is the following Markov chain based on the idea of splitting of the SDE flow (e.g., Milstein and Tretyakov, 2021):

$$\hat{X}^h_{n+1} = F_{X^h_n}\left(-\frac{h}{2}\nabla\phi(X^h_n)\right), \tag{3.13}$$

$$X^h_{n+1} = P_{\hat{X}^h_{n+1}}(h, \xi_{n+1}), \tag{3.14}$$

where F can be the exponential map or a retraction as described above, $\hat{X}^h_{n+1}$ denotes the intermediate step, and $P(h, \xi_{n+1}) : TM \to M$ and ξ_{n+1} are such that

$$\mathbb{E}\left(f\left(P_{\hat{X}^h_{n+1}}(h, \xi_{n+1})\right)\right) = f(\hat{X}^h_{n+1}) + \frac{h}{2}\Delta_M f(\hat{X}^h_{n+1}) + O(h^2) \tag{3.15}$$

for any $f \in C^{4,\epsilon}(M)$. The above relation can be satisfied, for example, if P is an exact simulation of Brownian motion from t_n to $t_n + h$, or if

$$P_{\hat{X}^h_{n+1}}(h, \xi_{n+1}) = F_{\hat{X}^h_{n+1}}\left(h^{1/2}g^{-1/2}(\hat{X}^h_{n+1})\xi_{n+1}\right). \tag{3.16}$$

The sampling method of Gatmiry and Vempala (2022) for Hessian manifolds can be considered as a special case of (3.13)-(3.14) with F being exponential map at X^h_n and P being exact simulation of Brownian motion starting at $\hat{X}^h_{n+1}$.

4. Error Bounds

In this section, first-order weak error bounds are given for the ensemble- and time-averaging estimators of $\mu_\phi(\varphi) = \int_M \varphi d\mu_\phi$, based, respectively, on a single and multiple independent trajectories of X^h_n obtained from (3.3) or (3.4). Proofs of the three theorems stated here are in the Appendix. We prove the theorems under the stated sufficient conditions which include smoothness assumptions on ϕ, φ and g. At the same time, the algorithms presented can be successfully used when, for example, φ is less smooth. In this respect we mention that in the Euclidean case (i.e, $M = \mathbb{R}^q$) weak convergence of the Euler scheme with order one was proved for φ being just measurable and bounded (Bally and Talay, 1995). Relaxing smoothness conditions for the theorems we prove here is a subject of future study.

Theorem 1 is proved for the Riemannian Langevin algorithm using a retraction (3.4). For clarity of exposition, Theorems 2 and 3 are proved for the algorithm (3.3) which uses the exponential map; proofs of these can be generalised to accommodate retractions in a manner similar to the proof of Theorem 1.

4.1 Ensemble-averaging estimator

Equation (2.11) implies that $\mu_\phi(\varphi)$ may be estimated upon using $\mathbb{E}(\varphi(X(T)))$ by choosing a sufficiently large T. This motivates the ensemble-averaging estimator $\hat{\mu}_{\phi,N}(\varphi)$ based on independent realizations $\{X^{(l),h}_N, l = 1, \ldots, L\}$ of X^h_n from the algorithm (3.4) up until the

final step N (recall that $T = Nh$), since

$$\hat{\mu}_{\phi,N}(\varphi) = \frac{1}{L} \sum_{l=1}^{L} \varphi(X_N^{(l),h}) \approx \mathbb{E}(\varphi(X_N^h)) \approx \mathbb{E}(\varphi(X(T))) \approx \mu_\phi(\varphi). \tag{4.1}$$

There are three sources of errors of $\hat{\mu}_{\phi,N}(\varphi)$: (i) proximity to the ergodic limit controlled by a choice of T; (ii) the numerical integration error controlled by the time step h; and (iii) the Monte Carlo error controlled by the number of Monte Carlo runs L. The following result shows how the weak upper bound on its bias depend on both h and T, and in a specific way on the class of functions φ. Note that the initial point x in (2.8) can be random, independent of the Wiener process $w(\cdot)$, and, consequently, X_0^h in the ensemble-averaging setting can be random, independent of ξ_n.

Theorem 1 (Bias of the ensemble averaging estimator) *Under Assumptions 1 and 2 and $\varphi \in C^{4,\epsilon}(M)$, the following bound holds for the the Riemannian Langevin algorithm with retraction (3.4) satisfying Assumption 3:*

$$|\mathbb{E}(\hat{\mu}_{\phi,N}(\varphi)) - \mu_\phi(\varphi)| \leq C\left(h + e^{-\lambda T}\right), \tag{4.2}$$

where C, λ are positive constants independent of h and T and the constant C linearly depends on $\|\varphi\|_{C^{4,\epsilon}(M)}$ but otherwise C, λ are independent of φ.

Variance of the estimator $\hat{\mu}_{\phi,N}(\varphi)$ is estimated in the usual way for sample means (see also Section 6):

$$\text{Var}(\hat{\mu}_{\phi,N}(\varphi)) = \frac{1}{L}\text{Var}(\varphi(X_N^h)) = \frac{1}{L}[\text{Var}(\varphi(X(T))) + O(h)] \tag{4.3}$$

$$= \frac{1}{L}[\mu_\phi(\varphi^2) - (\mu_\phi(\varphi))^2 + O(h + e^{-\lambda T})].$$

4.2 Time-averaging estimator

Since $X(t)$ is an ergodic process, we also have

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T \varphi(X(t))\mathrm{d}t = \mu_\phi(\varphi), \quad a.s., \tag{4.4}$$

which suggests that we can take $\frac{1}{T}\int_0^T \varphi(X(t))\mathrm{d}t$ as time-averaging estimator for $\mu_\phi(\varphi)$. Hence, the numerical time-averaging estimator is

$$\tilde{\mu}_{\phi,N}(\varphi) := \frac{1}{N} \sum_{n=0}^{N-1} \varphi(X_n^h). \tag{4.5}$$

In time-averaging estimation, we simulate a long trajectory and average φ at the discretized points collected along the long trajectory as can be seen in (4.5). There are again three errors associated with the estimator (4.5): (i) due to proximity to the ergodic limit controlled by a choice of T (i.e., by the choice of the number of steps N under fixed time step h), (ii) the numerical integration error controlled by the time step h, and (iii) due to variance of the estimator controlled by T.

19

Theorem 2 (Bias and mean-square error of the time-averaging estimator) *Under Assumptions 1 and 2 and $\varphi \in C^{2,\epsilon}(M)$, the following bounds hold for the Riemannian Langevin algorithm (3.3):*

$$|\mathbb{E}(\tilde{\mu}_{\phi,N}(\varphi)) - \mu_\phi(\varphi)| \leq C\left(h + \frac{1}{T}\right), \tag{4.6}$$

$$\mathbb{E}(\tilde{\mu}_{\phi,N}(\varphi) - \mu_\phi(\varphi))^2 \leq C\left(h^2 + \frac{1}{T}\right), \tag{4.7}$$

where $C > 0$ is independent of h and T and it linearly depends on $\|\varphi\|_{C^{2,\epsilon}(M)}$ but otherwise it is independent of φ.

We remark that the error C/T in (4.6)-(4.7) is due to the two sources of errors - (i) and (iii) as described before the theorem. They correspond to properties of the considered SDE (2.7). Further, it is not difficult to see that the estimates (4.6)-(4.7) imply that $\mathrm{Var}(\tilde{\mu}_{\phi,N}(\varphi)) \leq C(h^2 + 1/T)$, however one can prove under a mixing condition that $\mathrm{Var}(\tilde{\mu}_{\phi,N}(\varphi)) \leq C/T$ with $C > 0$ independent of h and T (see (Mattingly et al., 2010, Section 7) for more details when $M = \mathbb{T}^q$). For a comparison of computational costs between ensemble averaging and time averaging see (Tretyakov, 2025, Remark 3.9).

4.3 Bound on distance to invariant measure

The exponential map $TM \ni (x,v) \mapsto \exp_x(v) \in M$, and more generally retractions F_x, are continuous functions of $x \in M$. This and Assumption 2 imply that the recursion

$$X_{n+1}^h = \Phi(X_n^h, \xi_{n+1})$$

underlying the algorithm engenders a continuous map $\Phi : M \to M$ for every realization of ξ. As a result, $\{X_n^h\}$ arising from the proposed algorithm is a Feller chain: for a sequence $x_k \in M$ such that $\rho(x_k, x) \to 0$ as $k \to \infty$, $\mathbb{E}\varphi(\Phi(x_k, \xi)) \to \mathbb{E}\varphi(\Phi(x, \xi))$ as $k \to \infty$ for φ belonging to the class of bounded and continuous functions. A Feller chain evolving on a compact separable state space has a stationary measure (not necessarily unique) by the Krylov-Bogoliubov theorem (Da Prato and Zabczyk, 1996). Denote by μ_ϕ^h a stationary measure of the Markov chain X_n^h from (3.3). The following result provides an upper bound on an integral probability metric between μ_ϕ^h and μ_ϕ with respect to a class of test functions.

Theorem 3 (Distance to invariant measure)
Under the assumptions of Theorem 2, with

$$\mathrm{dist}(\mu, \nu) := \sup_{\varphi \in \mathcal{H}} \left| \int_M \varphi \mathrm{d}\mu - \int_M \varphi \mathrm{d}\nu \right|,$$

where $\mathcal{H} := \{\varphi \in C^{2,\epsilon}(M) \text{ and } \|\varphi\|_{C^{2,\epsilon}(M)} \leq 1\}$, the following bound holds:

$$\mathrm{dist}(\mu_\phi, \ \mu_\phi^h) \leq Ch, \tag{4.8}$$

with $C > 0$ being independent of h.

Remark 1 *Results similar to those in Theorems 1-3 can be proved for the two variants of the Riemannian Langevin algorithm discussed in Section 3.3. This is a consequence of the proof techniques used, which exploit links between the backward Kolmogorov and Poisson PDEs and the semigroup of the Langevin diffusion.*

4.4 Error extrapolation: increasing accuracy order and controlling bias

Analogous to the Euclidean case (Talay and Tubaro, 1990) (see also (Milstein and Tretyakov, 2021, Section 2.2.3)), it is possible to expand the errors of the bias of the estimators considered earlier in this section in powers of time increment h assuming additional smoothness of g, ϕ and φ, which is known as the Talay-Tubaro (or Richardson-Runge) extrapolation. For brevity, we skip technical details and merely highlight its practical importance. For definiteness, let us consider the bias of the ensemble-averaging estimator $\hat{\mu}_{\phi,N}(\varphi)$ with X_n^h obtained from (3.3), but it will become clear that the estimator $\tilde{\mu}_{\phi,N}(\varphi)$ may be treated analogously. Under appropriate smoothness assumptions, one can prove that

$$\lim_{N\to\infty} \mathbb{E}(\hat{\mu}_{\phi,N}(\varphi)) - \mu_\phi(\varphi) = C_1 h + \cdots + C_p h^p + O(h^{p+1}), \tag{4.9}$$

where the constants $C_1, \ldots, C_p$ are independent of h, T, and $p \geq 2$ is an integer (p can be arbitrarily large if g, ϕ and φ belong to $C^\infty(M)$).

Choose a sufficiently large T so that $e^{-\lambda T}$ in (4.2) is negligible, and take two time steps $h_1 = h$ and $h_2 = \alpha h$, $\alpha > 0$, $\alpha \neq 1$, with the corresponding number of steps being N_1 and N_2, respectively. Then, according to (4.9),

$$\begin{aligned} \mu_\phi(\varphi) &\approx \mathbb{E}(\hat{\mu}_{\phi,N_1}(\varphi)) + C_1 h_1 + O(h^2), \\ \mu_\phi(\varphi) &\approx \mathbb{E}(\hat{\mu}_{\phi,N_2}(\varphi)) + C_1 h_2 + O(h^2), \end{aligned} \tag{4.10}$$

whence

$$C_1 \approx -\frac{\mathbb{E}(\hat{\mu}_{\phi,N_2}(\varphi)) - \mathbb{E}(\hat{\mu}_{\phi,N_1}(\varphi))}{h_2 - h_1} + O(h). \tag{4.11}$$

By (4.10) and (4.11), we get the improved value with error $O(h^2)$:

$$\mu_\phi(\varphi)_{imp} := \mathbb{E}(\hat{\mu}_{\phi,N_1}(\varphi))\frac{h_2}{h_2 - h_1} - \mathbb{E}(\hat{\mu}_{\phi,N_2}(\varphi))\frac{h_1}{h_2 - h_1}. \tag{4.12}$$

Thus, the obtained method has an accuracy of order two in h. We can continue exploiting (4.9) and obtain a method of order three and so on. On the other hand, $C_1 h$ gives the leading term of the bias of $\hat{\mu}_{\phi,N_1}(\varphi)$, which can be used for error estimation in practice to assess the quality of the approximation. In Section 6.1, we apply the Talay-Tubaro extrapolation method to generate samples of the von-Mises Fisher distribution with bias of order 2.

4.5 Finite-time convergence

Although this paper is focused on the use of SDEs to sample from a given distribution on M, and hence on the estimators for ergodic limits from Section 4, for some applications it may also be of interest to discuss weak convergence of a Riemannian Langevin algorithm at a finite time T; this, for example, can be used to evaluate quantities related to transient behavior of diffusions on manifold and solving linear parabolic PDEs on manifolds by exploiting probabilistic representations (see (2.15)). It is straightforward to prove the following weak-sense convergence theorem along the lines of the proof of Theorem 1.

Theorem 4 (Finite-time convergence) *Under Assumptions 1 and 2 and $\varphi \in C^{4,\epsilon}(M)$, the Riemannian Langevin algorithm (3.3) for (2.7) is of weak order one, i.e.,*

$$|\mathbb{E}[\varphi(X_N) - \varphi(X(T))]| \leq Ch, \tag{4.13}$$

where $C > 0$ is a constant independent of h.

Remark 2 *Theorem 4 is more broadly applicable, and holds for weak-sense schemes similar to those in Section 3 for a general class of SDEs*

$$\mathrm{d}X(t) = a(t, X(t))\mathrm{d}t + \mathrm{d}B^M(t), \quad X(t_0) = x, \tag{4.14}$$

with a sufficiently smooth vector field $a(t,x)$ on $[t_0, T] \times M$.

Remark 3 *Results in this section demonstrate that the main techniques in the Euclidean setting on weak approximations and sampling carry over to the setting of compact manifolds M. In contrast, the situation with mean-square (strong) convergence is more delicate. For manifolds M where every point has an empty cut-locus, the Riemannian Langevin algorithm (3.3) with $\{\xi_n\}$ being independent standard Gaussian random variables has mean-square (strong) order $1/2$ (Bharath et al., 2025). Although this matches the result for Euclidean SDEs with multiplicative noise (e.g. Milstein and Tretyakov, 2021, Chapter 1), the proof in the manifold setting is substantially more complex. If a manifold contains points with non-empty cut-loci (e.g. positively curved manifolds) then we believe that mean-square schemes should be based on space-time random walks; see Milstein and Tretyakov (2021, Chapter 6), and also Milstein and Tretyakov (1999) for a mean-square approximation of Euclidean SDEs in bounded domains. This is in contrast to the weak-sense convergence, where the Riemannian Langevin algorithm on compact M has weak order 1, and is agnostic to the presence or absence of cut points.*

5. The Case of Non-compact Manifolds

In this section we discuss ingredients needed to transfer the error analysis of the preceding section from compact to non-compact M. As mentioned in Section 1, negative curvature encourages Brownian motion to drift away to infinity, and a uniform lower bound on the Ricci curvature is needed to address this. Additionally, since non-compactness is closely linked with non-positive curvature, when the Brownian motion has a drift component, as with an $\mathbb{R}^q$-valued Brownian motion, extra care is needed.

The condition (2.9) is sufficient to ensure stochastic completeness of a diffusion with generator $\mathcal{A}$ in (2.6) (Li, 1992); recall that it is automatically satisfied for compact M. Hence here we state (2.9) as the below assumption.

Assumption 4 *There exists a $\kappa > 0$ such that*

$$\mathrm{Ric}(v,v) + \mathrm{Hess}^{\phi}(v,v) \geq -\kappa g(v,v), \quad \text{for all } (x,v) \in TM.$$

Conditions in Assumption 4 may be further relaxed (Wang, 2013, Corollary 2.1.2), which, however, are harder to verify in practice. We note that Assumption 4 in the case of $M = \mathbb{R}^q$ resembles a one-sided Lipschitz condition which is sufficient to ensure regularity of SDEs in $\mathbb{R}^q$ as well as finiteness of moments of their solutions (Khasminskii, 2012).

Ergodicity, on the other hand, requires more. The *Bakry-Èmery criterion* (Bakry and Émery, 1985) demands the existence of a $\kappa > 0$ such that

$$\mathrm{Ric}(v, v) + \mathrm{Hess}^\phi(v, v) \geq 2\kappa g(v, v) \quad \text{for all} \ (x, v) \in TM, \tag{5.1}$$

which jointly controls both curvature of M and drift of the Brownian motion, and is more restrictive than (2.9). For example, when $M = \mathbb{R}^q$, the criterion (5.1) reduces to $\mathrm{Hess}^\phi \geq 2\kappa g$ for $\kappa > 0$, and implies that the density p is 2κ-strongly log-concave. The Bakry-Èmery criterion provides a sufficient condition for exponential ergodicity of the diffusion (2.7) on non-compact M, but is redundant for compact M. Recall that by construction, μ_ϕ is the invariant measure of the diffusion if it is ergodic.

A few intermediate conditions signpost the route to exponential ergodicity of the Langevin diffusion that satisfies the Bakry-Èmery criterion. The criterion ensures that the semigroup $\{P_t\}$ of operators corresponding to the diffusion satisfy the Log-Sobolev inequality with constant 2κ (Wang, 2009), which then implies that they satisfy the Poincaré inequality: $\mathrm{Var}_{\mu_\phi}(\varphi) \leq C\mathcal{E}(\varphi, \varphi)$ for a Dirichlet form $\mathcal{E}$ and constant C. As a consequence, one gets $|P_t|_{\mathrm{op}} \leq e^{-t/C}$, which implies exponential ergodicity.

The Bakry-Èmery criterion, however, is too stringent for our needs, as it precludes the possibility of sampling from a sizeable class of interesting probability measures on M. It is instead more natural to impose separate conditions on Ric and Hess^ϕ. The following assumption is sufficient for the semigroup of operators corresponding to the Langevin diffusion to satisfy the Log-Sobolev inequality (Wang, 2009, Theorem 1.1).

Assumption 5 *For constants $b, c > 0$,*

$$\mathrm{Ric} \geq (-c - b^2 \rho_o^2) g, \tag{5.2}$$

where $\rho_o = \rho(o, x)$, $o, x \in M$. The Hessian

$$\mathrm{Hess}^\phi \geq \delta g \tag{5.3}$$

outside of a compact set in M, where the constant δ satisfies $\delta > (1 + \sqrt{2})b\sqrt{q - 1} > 0$.

In the case $M = \mathbb{R}^q$ the assumption (5.3) is δ-strong convexity condition to be satisfied outside of a compact set in $\mathbb{R}^q$ (which means that potential function can be locally non-convex):

$$((x - y), \frac{1}{2}(\nabla\phi(x) - \nabla\phi(y))) \geq \delta|x - y|^2, \ \ \delta > 0,$$

which together with constant diffusion coefficient is sufficient for exponential ergodicity of the corresponding SDE in $\mathbb{R}^q$ (where Ric = 0) (Khasminskii, 2012; Roberts and Tweedie, 1996).

5.1 Examples

A few examples elucidate on Assumptions 4 and 5 for non-compact M.

Example 3 *Let M be the Poincaré half plane model $\mathbb{H}^q$ for the q-dimensional hyperbolic space with metric $g = \frac{1}{x_q^2} \sum_{i=1}^q \mathrm{d}x_i^2$, under which M has negative sectional curvature equal to -1 everywhere. From the Cartan-Hadamard theorem, it is diffeomorphic to $\mathbb{R}^q$. Consider the Riemannian-Gaussian distribution with density $p(x) \propto \exp(-\frac{\rho(x,o)^2}{2\sigma^2})$, with respect to dvol_g, where $o \in M$ is a fixed point and $\sigma > 0$; for brevity, let $\rho := \rho(x, o)$ be the distance function to the fixed o.*

The Hessian of the squared distance satisfies $\mathrm{Hess}^{\rho^2} \geq 2g$ (Greene and Wu, 2006), and since $\mathbb{H}^q$ is an Einstein manifold, its Ricci curvature $\mathrm{Ric} = -(q-1)g$ has a lower bound. Assumption 4 is satisfied since $\mathrm{Ric}+\mathrm{Hess}^\phi \geq (\frac{1}{\sigma^2}-q+1)g$, ensuring stochastic completeness. Note that the lower bound need not be positive as $\sigma > 0$ is unbounded, therefore the Bakry-Èmery criterion may be violated for some values of σ.

The lower bound on the Hessian means that we can restrict attention to a geodesic ball centered at o of finite radius to choose the values $c = (q-1)$, $b = ((1+\sqrt{2})\sigma^2\sqrt{q-1})^{-1/2}$ and $\delta = \frac{1}{\sigma^2}$ such that Assumption 5 is satisfied in tandem. Convergence exponentially fast to $p\,\mathrm{dvol}_g$ is thus assured for any $\sigma > 0$, including for values that violate the Bakry-Èmery criterion (5.1).

Example 4 *The space $M = \mathcal{P}_m$ consisting of $m \times m$ symmetric positive definite matrices with affine-invariant metric g in (6.6) is a $q = m(m+1)/2-$dimensional manifold. As with $\mathbb{H}^q$, the space $\mathcal{P}_m$ with metric g has a negative sectional curvature everywhere, which, however, unlike $\mathbb{H}^q$, is not constant. Assumptions 4 and 5 are verified for the following distributions, later used in Section 6.2.*

(i) *Let $\phi(X) = \frac{1}{2\sigma^2}\rho(X,O)^2$ in (2.3) for $\sigma > 0$ and fixed $O \in \mathcal{P}_m$. This choice of ϕ results in the Riemannian-Gaussian distribution (Section 6.2.1) on $\mathcal{P}_m$. It can be verified that $\mathrm{Ric} \geq -\frac{m}{4}g$, and by the Hessian comparison theorem $\mathrm{Hess}^{\rho^2} \geq 2g$ (e.g., Lewis, 2023, Section 4.4). For $\kappa = \left|\frac{1}{\sigma^2} - \frac{m}{4}\right|$, note that*

$$\mathrm{Ric} + \mathrm{Hess}^\phi \geq \left(\frac{1}{\sigma^2} - \frac{m}{4}\right)g,$$

and Assumption 4 is satisfied. For (5.2) in Assumption 5, we use the lower bound on the Ricci curvature with $c = \frac{m}{4}$. The value of b can be determined upon first calculating δ in (5.3). Since $\mathrm{Hess}^\phi \geq \frac{1}{\sigma^2}g$, we select $\delta = \frac{1}{\sigma^2}$ resulting in an interval $0 < b < ((1+\sqrt{2})\sigma^2\sqrt{m-1})^{-1/2}$ of possible values. Hence, Assumption 5 is satisfied. Note that the Bakry-Èmery criterion (5.1) is only satisfied for $\sigma < \sqrt{\frac{4}{m}}$.

(ii) *Let $\phi(X) = \rho(X,O)^4 - \rho(X,O)^2$ in (2.3) for fixed $O \in \mathcal{P}_m$. It will be shown in towards the end of this example that ϕ is geodesically non-convex as the Hessian is negative in some domain. To verify Assumption 4, note that a lower bound on the*

24

Ricci curvature is available from (i). To compute the Hessian, consider first a general function $f \in C^2(\mathcal{P}_m)$. By the chain rule

$$\mathrm{Hess}^{f(\rho^2)} = f''(\rho^2)\mathrm{d}\rho^2 \otimes \mathrm{d}\rho^2 + f'(\rho^2)\mathrm{Hess}^{\rho^2}.$$

Then setting $f(x) = x^2$ results in

$$\mathrm{Hess}^{\rho^4} = 2\mathrm{d}\rho^2 \otimes \mathrm{d}\rho^2 + 2\rho^2\mathrm{Hess}^{\rho^2},$$

such that Hessian of ϕ is

$$\mathrm{Hess}^{\phi} = D^2\rho^4 - D^2\rho^2 = 8\rho^2\mathrm{d}\rho \otimes \mathrm{d}\rho + (2\rho^2 - 1)\mathrm{Hess}^{\rho^2}$$

upon using the fact that $\mathrm{d}\rho^2 = 2\rho\mathrm{d}\rho$. Since $\mathrm{Hess}^{\rho^2} \geq 2g$ and $\mathrm{d}\rho \otimes \mathrm{d}\rho$ is a non-negative tensor:

$$\mathrm{Hess}^{\phi} \geq 2(2\rho^2 - 1)g.$$

Thus, $\mathrm{Ric} + \mathrm{Hess}^{\phi} \geq \left(2(2\rho^2 - 1) - \frac{m}{4}\right)g$ and Assumption 4 is satisfied for $\kappa = \frac{m+8}{4}$. Evidently, the Hessian is negative when $\rho < \frac{1}{\sqrt{2}}$, hence for the compact set we pick the closure of geodesic ball $\mathcal{B}_o(1)$ of radius 1 and centered at o. Outside of this ball, the Hessian is strictly positive, and choosing $\delta = 2$ ensures that (5.3) in Assumption 5 is satisfied. This implies that choosing b satisfying $0 < b < (2(1 + \sqrt{2})\sqrt{m - 1})^{-1/2}$ takes care of Assumption 5. Non-convexity of ϕ can now be gleaned from the fact that inside of $\mathcal{B}_o(1/\sqrt{2})$ the Hessian is negative. To see this, we look at the Hessian contracted with vector fields v such that $g(v, \nabla\rho) = 0$. Then

$$\mathrm{Hess}^{\phi}(v, v) = (2\rho^2 - 1)\mathrm{Hess}^{\rho^2}(v, v) \leq -2(1 - 2\rho^2)g(v, v) < 0.$$

5.2 Computational aspects

When $M = \mathbb{R}^q$ it is known that Euler-type schemes (including explicit ones) converge when coefficients of the SDE are globally Lipschitz (implying that the coefficients grow no faster than linearly at infinity) (Milstein and Tretyakov, 2021). It is quite straightforward to extend the analysis presented in this paper in the case of compact manifolds to non-compact case under the globally Lipschitz assumption on $\nabla\phi$ and $g^{-1/2}$. This extension will require to prove uniform bounds for moments of X_n^h.

However, distributions ϕ and manifolds (M, g) of practical interest are usually so that the globally Lipschitz assumption is violated. It is even so for Example 4(i) of the Riemannian-Gaussian distribution on $\mathcal{P}_m$.

It is known for SDEs on $M = \mathbb{R}^q$ that when growth of coefficients at infinity is faster than linear, *explicit* Euler schemes can diverge (Mattingly et al., 2002; Milstein and Tretyakov, 2021; Hutzenthaler et al., 2011). Reason for the divergence is exploding moments of the corresponding Markov chains (despite moments of the SDEs solution being bounded), typically due to a tiny number of exploding trajectories (Milstein and Tretyakov, 2005, 2021).

By rejecting exploding trajectories, one can safely use any method of weak approximation for a broad class of SDEs with nonglobally Lipschitz coefficients to compute the ensemble averaging estimator $\hat{\mu}_{\phi,N}(\varphi)$ of $\mu_{\phi}(\varphi)$. SDEs from this class need to merely satisfy

stochastic completeness and have sufficiently smooth coefficients, which SDEs of applicable interest typically do. The rejection technique relates to the second part of Assumption 5 (see (5.3)) in that one can view the condition as ignoring all sample paths of the diffusion that exit a sufficiently large geodesic ball $\mathcal{B}_o(R)$ for $R > 0$ around a point $o \in M$. Theoretical justification of this when $M = \mathbb{R}^q$ case is given by Milstein and Tretyakov (2005, 2007) (see also Milstein and Tretyakov (2021)), and it is possible to transfer the proof to the setting of non-compact manifolds M.

In Section 6.2 we provide numerical evidence of how the rejection technique can profitably be used by choosing a large enough closed geodesic ball, as in Assumption 5, on $\mathcal{P}_m$ using the proposed Riemannian Langevin Monte Carlo algorithm. In particular, no trajectory was rejected for $R = 2.7$ in the case of the Riemannian-Gaussian distribution (Section 6.2.1) and $R = 2$ for the non-convex potential (Section 6.2.2).

In $\mathbb{R}^q$, to address the difficulty with using the explicit Euler scheme for SDEs with non-globally Lipschitz coefficients in time-averaging estimators, one can use the Metropolis-adjusted Euler scheme (Bou-Rabee and Hairer, 2013). The Metropolis adjustment also either removes the bias or makes it exponentially small, however it bears substantial additional computational cost. A Metropolis-adjusted version of the Riemannian Langevin algorithm treated in this paper requires a special separate consideration.

Alternatives to the explicit Euler, that result in bounded moments and hence convergence, have been studied in depth for SDEs in $\mathbb{R}^q$ with non-globally Lipschitz coefficients. Examples include implicit methods (and in particular, the implicit Euler scheme) and explicit methods of tamed and balanced type (e.g., Mattingly et al., 2002; Hutzenthaler and Jentzen, 2015; Tretyakov and Zhang, 2013; Milstein and Tretyakov, 2021). Extensions of such methods to the non-compact manifold setting are possible and will be taken up elsewhere.

6. Numerical Illustrations

We carry out numerical experiments to verify the theoretical results from Section 4 for compact manifolds M, and demonstrate utility of the proposed algorithm for both compact and non-compact M. Section 6.1 considers sampling using retractions from the von-Mises distribution on the two-dimensional unit sphere $\mathbb{S}^2$. A sampling algorithm using extrinsic embedding coordinates was proposed by Wood (1994), while Fisher et al. (1993) proposed an exact method using intrinsic coordinates for dimension 2. A more recent development in Kurz and Hanebeck (2015) provides an extrinsic algorithm without rejection that is exact for even dimension spheres only.

Section 6.2 considers two distributions on the non-compact manifold of symmetric positive definite matrices: the Riemannian-Gaussian distribution with convex ϕ, and a distribution with non-convex potential. Versatility of the proposed algorithm to handle both convex and non-convex potential is demonstrated.

We focus on the ensemble-averaging estimator $\hat{\mu}_{\phi,N}(\varphi)$ defined in (4.1). To this end, we simulate a large number L of independent trajectories for a sufficiently long time $T = Nh$, and then verify that it enjoys first order of convergence in terms of time step h by comparing with the exact value of $\mu_\phi(\varphi)$ (obtained via numerical integration). Two types of errors are

reported: the numerical integration error

$$\texttt{err} = |\hat{\mu}_{\phi,N}(\varphi) - \mu_{\phi}(\varphi)|;$$

and, the Monte Carlo error

$$\texttt{MCerr} = \frac{1}{L-1}\left(\sum_{l=1}^{L} \varphi(X_N^{(l),h})^2 - \frac{1}{L}\left(\sum_{l=1}^{L} \varphi(X_N^{(l),h}) \right)^2 \right),$$

with corresponding 95% confidence interval

$$\hat{\mu}_{\phi,N}(\varphi) \pm 1.96\sqrt{\frac{\texttt{MCerr}}{L}}.$$

We note that in a similar manner, the theoretical results (see Theorem 2) for the time-averaging estimator $\hat{\varphi}_N$ (4.5) can also be verified.

Simulations were performed in R using a parallel architecture on 30 cores, using the packages `purrr` and `furrr` on a Supermicro 620U Linux RHEL8.8 server with 48 Intel Xeon (Ice Lake class) CPUs.

6.1 Retraction-based sampling from the von-Mises Fisher distribution on $\mathbb{S}^2$

The unit sphere $\mathbb{S}^2$ is a compact manifold with constant positive sectional curvature 1 everywhere, such that the cut locus of every point x is its antipode $-x$. This means that a single chart cannot cover $\mathbb{S}^2$ in order to implement the proposed algorithm. Charts can be defined by embedding $\mathbb{S}^2$ into $\mathbb{R}^3$, and this is commonly done using the inclusion map as the embedding such that $\mathbb{S}^2 = \{(x_1, x_2, x_3)^\intercal \in \mathbb{R}^3 : x_1^2 + x_2^2 + x_3^2 = 1\}$, and two stereographic projections from the north and south poles onto $\mathbb{R}^2$ are used to define two charts which cover $\mathbb{S}^2$.

We will instead consider spherical coordinates arising from the embedding of $\mathbb{S}^2$ into $\mathbb{R}^2$ given by $(r, \theta)^\intercal \mapsto (\sin r \cos \theta, \sin r \sin \theta, \cos r)^\intercal$, where r is the geodesic distance to the north pole, with $r \in [0, \pi], \theta \in [0, 2\pi]$. Our choice is motivated by the fact that expressions for the exponential and inverse-exponential maps are not available analytically in spherical coordinates with respect to the round metric g, and implementation of the proposed algorithm thus requires approximating the exponential map by numerically solving the geodesic equation. We thus consider a retraction of an appropriate order that preserves the order of errors derived in Section 4 (see Section 3.2), and we verify this numerically.

The Riemannian metric for the spherical coordinates at a point $x = (r, \theta)^\intercal$ is

$$g = \mathrm{d}r^2 + \sin^2 r \, \mathrm{d}\theta^2, \tag{6.1}$$

which is obtained by pulling back the Euclidean metric from $\mathbb{R}^3$. The inverse metric tensor assumes the form

$$G^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\sin^2 r} \end{pmatrix},$$

which is unbounded at $r = 0$ or $r = \pi$ or equivalently, at $(0, 0, 1)^\intercal$ or $(0, 0, -1)^\intercal$ in $\mathbb{R}^3$, and can lead to instability or loss of accuracy of the algorithm. The second chart is acquired by an orthogonal transformation of the coordinate axes obtained by permuting the

axes corresponding to the embedding $(r, \theta)^\intercal \mapsto (\cos r, \sin r \cos \theta, \sin r \sin \theta)^\intercal$ with the metric unchanged. The transition map between the charts is then

$$\psi(r, \theta) = (\arccos(\sin r \cos \theta), \arctan(\cot r \csc \theta))^\intercal, \tag{6.2}$$

which in $\mathbb{R}^3$ maps $(\sin r \cos \theta, \sin r \sin \theta, \cos r)^\intercal \mapsto (\cos r, \sin r \cos \theta, \sin r \sin \theta)^\intercal$; its inverse is

$$\psi^{-1}(\tilde{r}, \tilde{\theta}) = (\arccos(\sin \tilde{r} \sin \tilde{\theta}), \arctan(\tan \tilde{r} \cos \tilde{\theta}))^\intercal. \tag{6.3}$$

Numerically, we restrict the domain of the embeddings to $(\epsilon, \pi - \epsilon) \times [0, 2\pi]$ for some $\epsilon > 0$, whence the union of the two charts covers $\mathbb{S}^2$ as the singular points $r = 0$ and $r = \pi$ in chart 1 are mapped to $(\pi/2, \pi/2)$ and $(\pi/2, 3\pi/2)$, respectively, for any $\theta \in [0, 2\pi]$. We choose $\epsilon = 0.5$ in the simulations, which ensures that G^{-1} is bounded.

The centered von-Mises Fisher distribution μ_ϕ on $\mathbb{S}^2$ is the distribution

$$\mathrm{d}\mu_\phi \propto e^{\lambda \cos r} \mathrm{dvol}_g, \quad \text{for } \lambda > 0, \ r \in [0, \pi], \tag{6.4}$$

so that $\phi((r, \theta)^\intercal) = -\lambda \cos r$; performance of the algorithm for a non-centered version can be examined with obvious modifications. It was verified in Example 1 that the Langevin diffusion (2.7) with $\phi((r, \theta)^\intercal) = -\lambda \cos r$ is ergodic, and converges exponentially fast to its invariant measure μ_ϕ.

The geodesic equation (2.1) can be written as the following system of ODEs in local coordinates starting at $(r_0, \theta_0)^\intercal$:

$$\begin{aligned}
\dot{r} &= y, \ r(0) = r_0, \\
\dot{\theta} &= z, \ \ \theta(0) = \theta_0, \\
\dot{y} &= \sin r \cos r \ z^2, \ \ y(0) = v^1, \\
\dot{z} &= -2 \cot r \ yz, \ z(0) = v^2.
\end{aligned} \tag{6.5}$$

Since an exponential map in closed form is not available for this example, we use the retraction based on solving the above ODEs by a single step of size $\sqrt{h}$ of the standard $(1/6, 1/3, 1/3, 1/6)$ 4th-order Runge-Kutta method (RK4) as explained in Section 3.2.

For the testing purposes, we take $\varphi(x) = \sin r$ and compute, for $\lambda = 1$, the exact value $\mu_\phi(\varphi) = \sqrt{2} \frac{I_1(1)\Gamma(3/2)}{I_{1/2}(1)} \approx 0.7554024361$, where $I_n(\lambda)$ is the modified Bessel function of the first kind (Lewis, 2023, Section 4.1). The algorithm was terminated at time $T = 5$, sufficient to ensure that the error $|\mathbb{E}(\varphi(X(T))) - \mu_\phi(\varphi)|$ (i.e., the error due to proximity of the SDE's solution to the ergodic limit) is negligibly small. From Table 1 and the black line in Figure 1, we observe the first order convergence of Algorithm 1 as expected.

We also report results from an experiment with the modified Algorithm 1 in which the discrete random variables ξ distributed according to (3.2) are replaced with standard Gaussian random variables. The corresponding results given in Table 2 and the blue line in Figure 1 show first order convergence of Algorithm 1 with Gaussian random variables. Interestingly, the error is smaller in the case of Gaussian random variables than when the discrete random variables are used.

A further experiment was implemented in `Julia` using GPU kernels and run on an NVIDIA A100 GPU to compute the second order estimate of $\mu_\phi(\varphi)$ using the Talay-Tubaro

---

Algorithm 1 Sample from a von-Mises Fisher distribution on $\mathbb{S}^2$ using a retraction.

1. Input λ, h, N, X_0

2. Initialise the chain at $X_0 = (r_0, \theta_0)^\intercal$ according to the first chart and set chart counter $c = 1, \epsilon = 0.5$ and $n = 0$.

3. If $r_n < \epsilon$ or $r_n > \pi - \epsilon$:
   if $c = 1$, change the chart according to (6.2) and set $c = 2$;
   if $c = 2$, change the chart according to (6.3) and set $c = 1$.

4. Generate $\xi_{n+1} = (\xi^1_{n+1}, \xi^2_{n+1})^\intercal$, where ξ^i_{n+1} are independent, from (3.2).

5. Construct the tangent vector:
   if $c = 1$, set $v_{n+1} = -\frac{\lambda\sqrt{h}}{2} \sin r_n (1,0)^\intercal + g^{-1/2}(X_n^h)\xi_{n+1}$;
   if $c = 2$, set $v_{n+1} = \frac{\lambda\sqrt{h}}{2}(\cos r_n \sin\theta_n, \csc r_n \cos\theta_n)^\intercal + g^{-1/2}(X_n^h)\xi_{n+1}$.

6. Approximate the geodesic equation (6.5) with $(r_0, \theta_0)^\intercal = (r_n, \theta_n)^\intercal$, $(v^1, v^2)^\intercal = (v^1_{n+1}, v^2_{n+1})^\intercal$ by a single step of size $\sqrt{h}$ of RK4 to obtain $X_{n+1}^h = (r_{n+1}, \theta_{n+1})^\intercal$.

7. **If** $n + 1 = N$ then **stop** and if $c = 2$, change X_N^h to first chart according to (6.2), **else** put $n := n + 1$ and **return** to Step 3.

Table 1: Estimation error `err` and Monte Carlo error `MCerr` for sampling from a von-Mises Fisher distribution depending on the time step h and number of independent realisations L. The parameters are $\lambda = 1$, $X_0 = (r_0, \theta_0)^\intercal = (\pi/4, \pi/4)^\intercal$ and $T = 5$.

h	L	err	MCerr
0.2	10^6	0.0239	0.0004
0.1	10^6	0.0068	0.0004
0.05	10^6	0.0029	0.0005
0.025	10^7	0.0014	0.0001
0.0125	10^7	0.00065	0.00015
0.01	10^7	0.00048	0.00015

Table 2: Sampling from a von-Mises Fisher distribution (Algorithm 1 with Gaussian ξ). The parameters are as in Table 1.

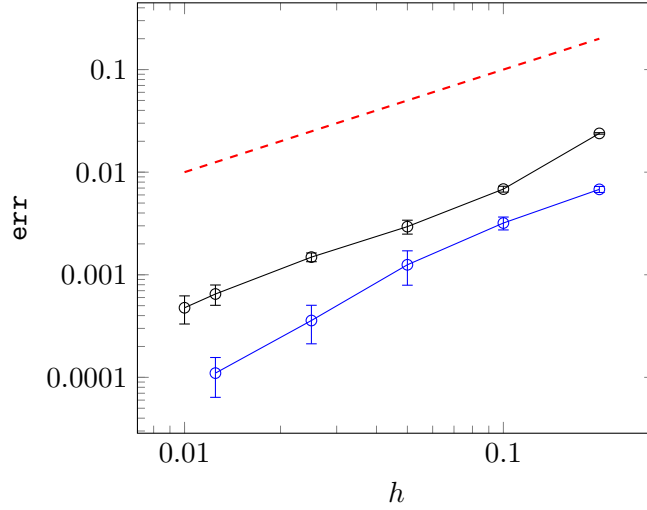| h	L	err	MCerr
0.2	10^6	0.0068	0.0005
0.1	10^6	0.0032	0.0005
0.05	10^6	0.0013	0.0005
0.025	10^7	0.00036	0.00015
0.0125	10^8	0.00011	0.00005

Figure 1: Sampling from a von-Mises Fisher distribution. A log-log plot of the error `err` against h. The black (blue) line corresponds to the Algorithm 1 with ξ_n distributed according to (3.1) (standard Gaussian distribution). Error bars correspond to Monte Carlo error `MCerr`. The reference red line has gradient 1. The parameters are as in Table 1.

extrapolation method detailed in Section 4.4. For the initial first-order estimates, the experiment was repeated with the time steps $h = 0.1, 0.08, 0.0625, 0.05$ and with parameters $T = 10$ and $L = 5 \times 10^9$ ($N_1 = 100$, $N_2 = 125$, $N_3 = 160$, $N_4 = 200$). The respective first-order estimates of $\mu_\phi(\varphi)$ were $\hat{\mu}_{\phi,N_1}(\varphi) = 0.762017 \pm 0.000006$ (`err` $\doteq 0.006615$), $\hat{\mu}_{\phi,N_2} = 0.760524 \pm 0.00006$ (`err` $\doteq 0.005122$), $\hat{\mu}_{\phi,N_3}(\varphi) = 0.759317 \pm 0.000006$ (`err`$\doteq 0.003915$) and $\hat{\mu}_{\phi,N_4}(\varphi) = 0.758523 \pm 0.000006$ (`err` $\doteq 0.003121$). Then, using (4.12) and the above data, the three second order estimates of $\mu_\phi(\varphi)$ are: $\hat{\mu}_{\phi,imp}(\varphi) = 0.754552 \pm 0.000041$ (`err` $\doteq 0.00085$) based on $\hat{\mu}_{\phi,N_1}(\varphi)$ and $\hat{\mu}_{\phi,N_2}(\varphi)$; $\hat{\mu}_{\phi,imp}(\varphi) = 0.755006 \pm 0.000037$ (`err` $\doteq 0.000396$) based on $\hat{\mu}_{\phi,N_2}(\varphi)$ and $\hat{\mu}_{\phi,N_3}(\varphi)$; and $\hat{\mu}_{\phi,imp}(\varphi) = 0.755348 \pm 0.000041$ (`err` $\doteq 0.000054$) based on $\hat{\mu}_{\phi,N_3}(\varphi)$ and $\hat{\mu}_{\phi,N_4}(\varphi)$. We see that the Talay-Tubaro extrapolation improves the accuracy as expected. The combined computation time of this experiment was about 3000 seconds thanks to an effective use of GPU.

6.2 Sampling on the manifold of SPD matrices

We consider the non-compact manifold of symmetric positive definite (SPD) matrices, and demonstrate numerically that the proposed Riemannian Langevin algorithm results in the desired error bounds when sampling from two distributions which satisfy the assumptions discussed in Section 5. In contrast to the unit sphere in Section 6.1, the algorithm is implemented using a closed form expression for the exponential map.

Consider the general linear group $GL(m)$ consisting of the set of $m \times m$ real invertible matrices equipped with the group operation of matrix multiplication. Denote by $M = \mathcal{P}_m$, the manifold of $m \times m$ symmetric positive definite matrices $\mathcal{P}_m = \{X \in GL(m) : X^\mathsf{T} = X, y^\mathsf{T} X y > 0, y \in \mathbb{R}^m \setminus \{0\}\}$, an open cone of dimension $q = m(m+1)/2$.

The tangent space $T_X \mathcal{P}_m = \{X\} \times \mathcal{S}_m$, where $\mathcal{S}_m$ is the set of $m \times m$ real symmetric matrices. Various choices of metric are available on $\mathcal{P}_m$ (Pennec et al., 2006), and we choose the affine invariant metric

$$g(U, V)_X = \text{Tr}(X^{-1}UX^{-1}V), \quad U, V \in T_X \mathcal{P}_m, \tag{6.6}$$

which is invariant under the transformation $X \mapsto A^{-1}XA$, for every $A \in GL(m)$. The metric is compatible with the transitive action of $GL(m)$ on $\mathcal{P}_m$ in that between any pair $X_1, X_2 \in \mathcal{P}_m$ there exists an $A \in GL(m)$ such that $X_2 = A^{-1}X_1A$, which makes $\mathcal{P}_m$ a homogeneous space of $GL(m)$.

Affine invariance of the metric ensures that the Riemannian distance ρ and volume form dvol_g will also be invariant under affine transformations. The metric g gives $\mathcal{P}_m$ a Riemannian structure with non-positive sectional curvature, and $\mathcal{P}_m$ is a non-compact, geodesically complete metric space by the Hopf-Rinow theorem (e.g., Jost, 2008). Moreover, by the Cartan-Hadamard theorem (e.g., Kobayashi and Nomizu, 1969), the inverse exponential map is globally defined, and the cut locus of every point is empty, i.e., every two points in $\mathcal{P}_m$ have a unique geodesic connecting them. This means that only a single chart is needed to cover $\mathcal{P}_m$.

The volume form at a point $X \in \mathcal{P}_m$ with affine-invariant metric g is

$$\text{dvol}_g = \det(X)^{(m+1)/2}\text{d}x_1 \ldots \text{d}x_{m(m+1)/2},$$

where $x_1, \ldots, x_{m(m+1)/2}$ are the upper triangular elements of X. For $X \in \mathcal{P}_m$ and $S \in \mathcal{S}_m$, the exponential map has the closed form

$$\exp_X(tS) = X^{1/2}\text{Exp}(tX^{-1/2}SX^{-1/2})X^{1/2}, \tag{6.7}$$

where Exp is the matrix exponential. A geodesic that connects two points $X_1, X_2 \in \mathcal{P}_m$ can be parameterised as

$$\gamma(t) = X_1^{1/2}\text{Exp}(t\text{Log}(X_1^{-1/2}X_2X_1^{-1/2}))X_1^{1/2}$$

so that $\gamma(0) = X_1$ and $\gamma(1) = X_2$, where Log is the matrix logarithm defined globally on $\mathcal{P}_m$. The distance between two points X_1 and X_2 is

$$\rho(X_1, X_2) = \sqrt{\sum_{i=1}^{m}(\log r_i)^2},$$

where $\{r_i\}_{i=1}^m$ are the eigenvalues of $X_1^{-1}X_2$. The inverse exponential map can also be written in the closed form:

$$\exp_{X_1}^{-1}(X_2) = \dot{\gamma}(0) = X_1^{1/2}\text{Log}(X_1^{-1/2}X_2X_1^{-1/2})X_1^{1/2}. \tag{6.8}$$

For efficient use of the coordinates, we consider the vectorization and half-vectorization maps. With $\text{vec} : \mathcal{S}_m \to \mathbb{R}^{m^2}$ as the vectorization map taking a symmetric matrix to a column vector, consider the half-vectorization map

$$\text{hvec} : \mathcal{S}_m \to \mathbb{R}^q, \quad \text{hvec}(x) := B\text{vec}(x),$$

where the matrix

$$B := \sum_{i \geq j} (u_{ij} \otimes e_j^{\mathsf{T}} \otimes e_i^{\mathsf{T}}) \in \mathbb{R}^{q \times m^2}$$

picks out the lower triangular part of the vectorization, and u_{ij} is a q-dimensional unit vector with 1 in position $(j-1)m + i - j(j-1)/2$ and 0 elsewhere, and e_i is the standard basis of $\mathbb{R}^{m^2}$. Its inverse

$$\mathbb{R}^q \ni x \mapsto \mathsf{hvec}^{-1}(x) := (\mathsf{hvec}(I_m)^{\mathsf{T}} \otimes I_m)(I_m \otimes x) \in \mathbb{R}^{m \times m}$$

exists through the Moore-Penrose inverse of A.

For the experiments, we fix $m = 3$ and consider $\mathcal{P}_3$ with global coordinates $x = (x_1, x_2, x_3, x_4, x_5, x_6)$ such that

$$X = \begin{pmatrix} x_1 & x_4 & x_6 \\ x_4 & x_2 & x_5 \\ x_6 & x_5 & x_3 \end{pmatrix}.$$

The metric tensor G is expressed as $G = \mathrm{d}x^{\mathsf{T}}(X^{-1} \otimes X^{-1})\mathrm{d}x$, where $\mathrm{d}x := \mathsf{hvec}(\mathrm{d}X)$ with $\mathrm{d}X = (\mathrm{d}x_i)$. The inverse metric G^{-1} can be concisely expressed (e.g., Moakher and Zéraï, 2011):

$$G^{-1}(X) = \begin{pmatrix} x_1^2 & x_4^2 & x_6^2 & x_1 x_4 & x_4 x_6 & x_1 x_6 \\ x_4^2 & x_2^2 & x_5^2 & x_2 x_4 & x_2 x_5 & x_4 x_5 \\ x_6^2 & x_5^2 & x_3^2 & x_5 x_6 & x_3 x_5 & x_3 x_6 \\ x_1 x_4 & x_2 x_4 & x_5 x_6 & \frac{1}{2}\left(x_1 x_2 + x_4^2\right) & \frac{1}{2}(x_2 x_6 + x_4 x_5) & \frac{1}{2}(x_1 x_5 + x_4 x_6) \\ x_4 x_6 & x_2 x_5 & x_3 x_5 & \frac{1}{2}(x_2 x_6 + x_4 x_5) & \frac{1}{2}\left(x_2 x_3 + x_5^2\right) & \frac{1}{2}(x_3 x_4 + x_5 x_6) \\ x_1 x_6 & x_4 x_5 & x_3 x_6 & \frac{1}{2}(x_1 x_5 + x_4 x_6) & \frac{1}{2}(x_3 x_4 + x_5 x_6) & \frac{1}{2}\left(x_1 x_3 + x_6^2\right) \end{pmatrix},$$

which must be written as a column vector in $\mathbb{R}^6$ using the hvec^{-1} map for use in the algorithms.

6.2.1 Riemannian-Gaussian distribution

The Riemannian-Gaussian distribution can be viewed as the generalisation of the isotropic normal distribution on $\mathbb{R}^q$ to $\mathcal{P}_m$, where the Euclidean distance is replaced by the Riemannian distance in the probability density function:

$$\mathrm{d}\mu_\phi(X) = \frac{1}{Z_m(\sigma)} e^{-\phi(X)} \mathrm{dvol}_g, \ X \in \mathcal{P}_m,$$

where

$$\phi(X) = \frac{1}{2\sigma^2} \rho(X, O)^2 \tag{6.9}$$

for a fixed $O \in \mathcal{P}_m$ and parameter $\sigma > 0$. The function $x \mapsto \phi(x)$ is strictly convex, since, by the Hessian comparison theorem (Cheeger and Ebin, 1975) the map $x \mapsto \rho(x, \cdot)^2$ is geodesically convex on a manifold with everywhere non-positive sectional curvature.

The normalisation constant $Z_m(\sigma)$ can be evaluated by employing spectral decomposition $X = Q^\mathsf{T}\mathrm{diag}(e^{\lambda_1}, \ldots, e^{\lambda_m})Q$, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_m > 0$, and computing the following integral

$$c_m \int_{\mathbb{R}^m} e^{-\frac{1}{2\sigma^2}[\lambda_1^2 + \cdots + \lambda_m^2]} \prod_{i<j} \sinh\left(\frac{|\lambda_i - \lambda_j|}{2}\right) \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_m, \tag{6.10}$$

where $c_m = \frac{1}{m!} \frac{\pi^{m^2/2}}{\Gamma_m(m/2)} 8^{m(m-1)/4}$ and Γ_m is the multivariate gamma function (Said et al., 2017); the constant c_m is part of the volume form dvol_g, and when $m = 3$, $c_3 = \frac{16\sqrt{2}\pi^2}{3}$. Without loss of generality we set $O = I_m$, since $\mathcal{P}_m$ is a homogeneous space. For convenience, we use $\rho_I = \rho_I(X) := \rho(X, I_m)$.

Table 3: Estimation and Monte Carlo errors for the Riemannian—Gaussian distribution. The parameters are $O = I_3$, $\sigma = \frac{1}{\sqrt{2}}$, $X_0 = \mathsf{hvec}^{-1}((2, 4, 2, 1, 1, 0)^\mathsf{T})$ and $T = 10$.

h	L	err	MCerr
0.2	10^6	0.148	0.008
0.1	10^6	0.078	0.008
0.05	10^7	0.035	0.002
0.025	10^7	0.017	0.002

To our knowledge, currently, only a rejection-based algorithm for sampling from Riemannian-Gaussian on $\mathcal{P}_m$, which relies on sampling the eigenvalues, is available in the literature (Said et al., 2017). Our Riemannian Langevin algorithm given in Algorithm 2 does not contain a rejection step; however, a rejection step can in principle be incorporated if needed (for example, to be used as a proposal distribution within MCMC).

The Riemannian-Gaussian distribution satisfies Assumptions 4 and 5, as seen in Example 4 (i). This ensures that the Langevin diffusion converges to its invariant measure exponentially fast. From (6.8) the gradient of ϕ is

$$\nabla\phi(X) = -\frac{\exp_X^{-1}(I)}{\sigma^2} = \frac{X^{1/2}\mathrm{Log}(X)X^{1/2}}{\sigma^2},$$

which grows at infinity as $x \log x$; also note that $\nabla\phi(I) = 0$.

The function $\varphi : \mathcal{P}_3 \to \mathbb{R}$ we consider to assess accuracy is $\varphi(X) = \det(X)$. The value of $\mu_\phi(\varphi) \approx 2.11699998$, evaluated with accuracy of order of 10^{-8} using the function `NIntegrate` in Mathematica. Table 3 and Figure 2 demonstrate that Algorithm 2 exhibits the first order convergence.

6.2.2 DISTRIBUTION WITH NON-CONVEX POTENTIAL

The second part of Assumption 5 (see (5.3)) makes it evident that we need not have log-concavity of the density function to guarantee ergodicity of the SDE. We verify this and demonstrate generality of our algorithm by considering the following distribution with a

---

Algorithm 2 Algorithm to sample from Riemannian-Gaussian distribution on $\mathcal{P}_3$ with $O = I_3$, $\sigma = \frac{1}{\sqrt{2}}$ and $X_0 = \mathsf{hvec}^{-1}((2, 4, 2, 1, 1, 0)^\intercal)$.

---

1. Input h, N

2. Initialise the chain at X_0 and set $k = 0$.

3. Generate $\xi_k = (\xi_k^1, \xi_k^2, \xi_k^3, \xi_k^4, \xi_k^5, \xi_k^6)^\intercal$, where, for every k, ξ_k^i are independent from (3.2).

4. Compute the matrix $M_k = -\frac{h}{2\sigma^2}\mathrm{Log}(X_k^h) + \sqrt{h}(X_k^h)^{-1/2}\mathsf{hvec}^{-1}(G^{-1/2}(X_k^h)\xi_k)(X_k^h)^{-1/2}$.

5. Set $X_{k+1}^h = (X_k^h)^{1/2}\mathrm{Exp}(M_k)(X_k^h)^{1/2}$.

6. **If** $k + 1 = N$ then **stop**, **else** put $k := k + 1$ and **return** to Step 3.
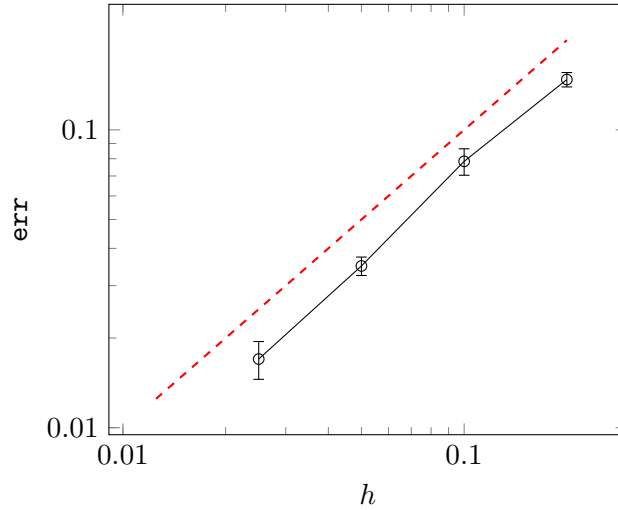
Figure 2: Sampling from the Riemannian—Gaussian distribution on $\mathcal{P}_3$: log-log plot of the estimation error `err` against h. Error bars correspond to Monte Carlo error `MCerr`. The reference red line has gradient 1. The parameters for the Riemannian-Gaussian are as in Table 3.

non-convex potential ϕ:

$$\mathrm{d}\mu_\phi = \frac{1}{Z_m}e^{-\phi(X)}\mathrm{dvol}_g, \quad X \in \mathcal{P}_m,$$

where

$$\phi(X) = \rho(X,O)^4 - \rho(X,O)^2,$$

with the normalisation constant Z_m. This distribution is referred to as one with a double-well potential. Similar to the Riemannian-Gaussian distribution, the constant Z_m is obtained by evaluating

$$c_m \int_{\mathbb{R}^m} e^{-\left([\lambda_1^2+\cdots+\lambda_m^2]^2 - [\lambda_1^2+\cdots+\lambda_m^2]\right)} \prod_{i<j} \sinh\left(\frac{|\lambda_i - \lambda_j|}{2}\right) \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_m,$$

where c_m is as in (6.10).

Example 4 (ii) verified that $\phi(X)$ satisfied Assumptions 4 and 5. The corresponding Langevin diffusion (2.7) converges exponentially fast to the invariant distribution μ_ϕ. The gradient takes the form

$$\nabla\phi(X) = -4\rho(X,I_m)^2 \exp_X^{-1}(I_m) + 2\exp_X^{-1}(I_m) = (4\rho(X,I_m)^2 - 2)X^{1/2}\mathrm{Log}(X)X^{1/2}.$$

Algorithm 3 Algorithm to sample from double-well potential distribution on $\mathcal{P}_3$ with $O = I_3$ and $X_0 = \mathsf{hvec}^{-1}((2,4,2,1,1,0)^\mathsf{T})$.

---

1. Input h, N.

2. Initialise the chain at X_0 and set $k = 0$.

3. Generate $\xi_k = (\xi_k^1, \xi_k^2, \xi_k^3, \xi_k^4, \xi_k^5, \xi_k^6)^\mathsf{T}$, where, for every k, ξ_k^i are independent from (3.2).

4. Compute the matrix

$$M_k = h(1 - 2\rho(X_k^h, I)^2)\mathrm{Log}(X_k^h) + \sqrt{h}(X_k^h)^{-1/2}\mathsf{hvec}^{-1}(G^{-1/2}(X_k^h)\xi_k)(X_k^h)^{-1/2}.$$

5. Set $X_{k+1}^h = (X_k^h)^{1/2}\mathrm{Exp}(M_k)(X_k^h)^{1/2}$.

6. **If** $k + 1 = N$ then **stop**, **else** put $k := k+1$ and **return** to Step 3.

We choose $\varphi(X) = 1/(1 + \mathrm{tr}(X))$ and $\mu_\phi(\varphi) \approx 0.2204801571878534$, evaluated with accuracy of order of 10^{-8} using the function `NIntegrate` in Mathematica. Table 4 and Figure 3 show that Algorithm 3 exhibits the first order convergence.

7. Discussion and Concluding Remarks

In this paper we have studied how the intrinsic Langevin diffusion can be used for sampling on Riemannian manifolds. To this end we have analyzed and provided error estimates for intrinsic Riemannian Langevin algorithms including those using retractions when the exponential map is not available in closed form. The algorithm ensures that the corresponding Markov chain moves on the manifold M without requiring a projection, i.e. it preserves the geometrical features of the corresponding stochastic differential equation.

Table 4: The double-well potential distribution. The parameters are $O = I_m$, $X_0 = \mathsf{hvec}^{-1}((2, 4, 2, 1, 1, 0)^\intercal)$ and $T = 5$.

h	L	err	MCerr
0.2	10^5	0.00537	0.0003
0.1	10^5	0.00162	0.0003
0.05	10^7	0.000605	0.00003
0.025	10^7	0.000258	0.00003
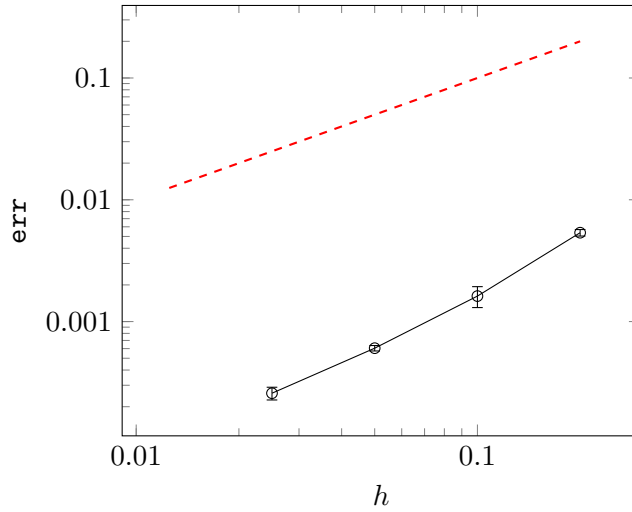
Figure 3: Sampling from a distribution with non-convex 'double-well' potential on $\mathcal{P}_3$: log-log plot of the estimation error err against h. Error bars correspond to Monte Carlo error MCerr. The reference red line has gradient 1. The parameters are as in Table 4.

The main focus is on sampling from, and computing expectation of functions of, a target distribution. As a consequence, we demonstrate how the concept of weak approximations for SDEs in $\mathbb{R}^q$ and the flat torus $\mathbb{T}^q$, and the corresponding proof techniques in deriving weak error bounds, can be transferred to the case of compact Riemannian manifolds via the covariant Taylor expansion. The PDE-based proof technique enables obtaining weak error bounds that match the optimal bounds in $\mathbb{R}^q$ and flat manifolds; these are superior to the weak error bound currently available in the literature, which follow by first deriving strong approximation (e.g., convergence in expected squared-distance) error bounds. The price we pay for our approach is in the absence of explicit dimension-dependent constants in the derived bounds. The Talay-Tubaro extrapolation in Section 4.4 may be used to extract more information on the constants.

The approach used here to deriving weak error bounds opens the door to carry over the extensive research (conducted by statistical, machine learning, and the numerical analysis communities) on approximation of SDEs in Euclidean spaces to manifolds. For example, sampling from a distribution with compact support on a non-compact manifold, or from a manifold with boundary, are of considerable interest given their practical relevance. These problems are yet to be considered within current literature; only recently has the former been considered even for $\mathbb{R}^q$ using a reflected diffusion (Leimkuhler et al., 2023). It is plausible that an intrinsic reflected diffusion can be used to carry out a similar program on manifolds, and this is the subject of our future work.

In this paper we have only considered explicit Euler-type schemes. In a similar fashion, we can construct and analyze implicit schemes, second-order sampling algorithms, etc. for Langevin diffusion on manifolds using the arsenal of numerical methods developed in the Euclidean space (see e.g. (Milstein and Tretyakov, 2021)), which potentially can lead to more efficient sampling methods.

The proposed algorithm and its analysis are also useful for applications in other research areas, including molecular dynamics, finance, and optimisation. An important example in financial engineering involves multi-factor stochastic volatility models widely used in option pricing and risk analysis, where stochastic modelling of correlation/covariance between factors entering the asset price and volatility processes are of practical value. There is growing interest (Fonseca et al., 2008; Alfonsi et al., 2019) in using a Wishart process (Bru, 1991) to model the time-changing covariance structure, which is an example of stochastic process assuming values in the manifold of positive (semi-)definite matrices. The examples considered in Section 6.2 with the intrinsic approach to modelling stochastic dynamics of positive definite matrices provides a useful, more general class of processes to be used within this context.

Acknowledgements

Appendix A.

The covariant Taylor expansion plays a major role in the proofs of theorems, and we review it here. Let $f \in C^4(M)$ taking values in $\mathbb{R}$ and the curve $c(s)$, $s \in \mathbb{R}$ on M, be so that $c(0) = x$ and $\dot{c}(0) = V$. Introduce $F(s) := f(c(s)) : \mathbb{R} \to \mathbb{R}$ for which the Taylor expansion can be written as

$$F(s) = F(0) + s\frac{\mathrm{d}}{\mathrm{d}s}F(0) + \frac{s^2}{2}\frac{\mathrm{d}2}{\mathrm{d}s^2}F(0) + \frac{s^3}{6}\frac{\mathrm{d}3}{\mathrm{d}s^3}F(0) + \frac{s^4}{24}\frac{\mathrm{d}4}{\mathrm{d}s^4}F(\alpha), \tag{A.1}$$

where $\alpha \in (0, s)$. We examine each derivative in turn. The first derivative is

$$\frac{\mathrm{d}}{\mathrm{d}s}F(s) = \frac{\mathrm{d}}{\mathrm{d}s}f(c(s)) = \mathrm{d}f(c(s))(\dot{c}(s)). \tag{A.2}$$

For the second derivative, the geodesic curvature of c is accounted for

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}s^2}F(s) &= \frac{\mathrm{d}^2}{\mathrm{d}s^2}f(c(s)) \\
&= \frac{\mathrm{d}}{\mathrm{d}s}\mathrm{d}f(c(s))(\dot{c}(s)) \\
&= D\mathrm{d}f(c(s))(\dot{c}(s), \dot{c}(s)) + \mathrm{d}f(c(s))(D_{\dot{c}(s)}\dot{c}(s)).
\end{aligned} \tag{A.3}$$

Computing the third derivative, we get

$$\begin{aligned}
\frac{\mathrm{d}^3}{\mathrm{d}s^3}F(s) &= \frac{\mathrm{d}}{\mathrm{d}s}\left(D\mathrm{d}f(c(s))(\dot{c}(s), \dot{c}(s)) + \mathrm{d}f(c(s))(D_{\dot{c}(s)}\dot{c}(s)) \right) \\
&= D^2\mathrm{d}f(c(s))(\dot{c}(s), \dot{c}(s), \dot{c}(s)) + 2D\mathrm{d}f(c(s))(\dot{c}(s), D_{\dot{c}(s)}\dot{c}(s)) \\
&\quad + D\mathrm{d}f(c(s))(\dot{c}(s), D_{\dot{c}(s)}\dot{c}(s)) + \mathrm{d}f(c(s))(D_{\dot{c}(s)}D_{\dot{c}(s)}\dot{c}(s)) \\
&= D^2\mathrm{d}f(c(s))(\dot{c}(s), \dot{c}(s), \dot{c}(s)) + 3D\mathrm{d}f(c(s))(\dot{c}(s), D_{\dot{c}(s)}\dot{c}(s)) \\
&\quad + \mathrm{d}f(c(s))(D_{\dot{c}(s)}D_{\dot{c}(s)}\dot{c}(s)),
\end{aligned} \tag{A.4}$$

where we have utilized the fact that second covariant derivative of functions is symmetric $(0, 2)$−tensor. Evaluating (A.2), (A.3), (A.4) at $s = 0$ and substituting them into (A.1), we obtain the following Taylor formula of f along the curve c:

$$\begin{aligned}
f(c(s)) &= f(x) + s\mathrm{d}f(x)(V) + \frac{s^2}{2}(D\mathrm{d}f(x)(V, V) + \mathrm{d}f(x)(D_V V)) \\
&\quad + \frac{s^3}{6}(D^3 f(x)(V, V, V) + 3D^2 f(x)(V, D_V V) + \mathrm{d}f(x)(D_V^2 V)) \\
&\quad + \frac{s^4}{24}\frac{\mathrm{d}4}{\mathrm{d}s^4}f(c(\alpha)),
\end{aligned} \tag{A.5}$$

where we have abbreviated $D_V^2 = D_V D_V$.

Appendix B. Proofs of Theorems

For brevity, we denote $u(t, x)$ as u, $\frac{\partial u}{\partial t}(t, x)$ as $\frac{\partial u}{\partial t}$, $\nabla u(t, x)$ as ∇u, $Du(t, x)$ as Du, $D^k u(t, x)$ as $D^k u$, $\nabla \phi(x)$ as $\nabla \phi$, $g^{-1/2}(x)$ as $g^{-1/2}$ and $u(t_n, X_n^h)$ as u_n, $n = 0, \ldots, N-1$, where $u(t, x)$ is the solution of (2.12) and $\phi(x)$ is from (2.7). We need the following lemma on one-step error of the algorithm (3.4) to prove Theorem 1.

Lemma 1 (One-step error) *Let Assumptions 1 and 2 and $\varphi \in C^{4,\epsilon}(M)$ hold. Given $x \in M$, let X_1^h be computed according to the following formula:*

$$X_1^h = F_x\left(-\frac{h}{2}\nabla\phi(x) + h^{1/2}g^{-1/2}(x)\xi\right), \tag{B.1}$$

where F_x is a retraction, and curve $c(s)$ corresponding to this retraction satisfies Assumption 3 with $c(0) = x$, $\dot{c}(0) = V := -\frac{h^{1/2}}{2}\nabla\phi + g^{-1/2}\xi$ and hence $F_x(V) = c(\sqrt{h})$. Then

$$\mathbb{E}[u(t + h, X_1^h) - u(t, x)] \leq Ch^2 e^{-\lambda(T-t)}, \tag{B.2}$$

where $C > 0$ is independent of T and h and it linearly depends on $\|\varphi\|_{C^{4,\epsilon}(M)}$ but otherwise is independent of φ.

Proof Applying the Taylor expansion to $u(t + h, c(s))$ around t, we obtain

$$u(t+h, c(s)) = u(t, c(s)) + h\frac{\partial u}{\partial t}(t, c(s)) + \frac{h^2}{2}\frac{\partial^2 u}{\partial t^2}(t(\alpha_1), c(s)), \tag{B.3}$$

where $t(\alpha_1) := t + \alpha_1 h \in (t, t+h)$ since $\alpha_1 \in (0, 1)$.

Applying the covariant Taylor formula (A.5) to the function $u(t, c(s))$ along the curve c, we get

$$u(t, c(s)) = u + s\mathrm{d}u(V) + \frac{s^2}{2}(D\mathrm{d}u(V, V) + \mathrm{d}u(D_V V))$$

$$+ \frac{s^3}{6}(D^3 u(V, V, V) + 3D^2 u(V, D_V V) + \mathrm{d}u(D_V^2 V))$$

$$+ \frac{s^4}{24}\frac{\mathrm{d}^4}{\mathrm{d}s^4}u(t, c(\alpha_2)), \tag{B.4}$$

where $\alpha_2 \in (0, \sqrt{h})$. Similarly we obtain

$$\frac{\partial u}{\partial t}(t, c(s)) = \frac{\partial u}{\partial t} + sD\frac{\partial u}{\partial t}(V)$$

$$+ \frac{s^2}{2}\left(D^2\frac{\partial u}{\partial t}(t, c(\alpha_3))(\dot{c}(\alpha_3), \dot{c}(\alpha_3)) + \mathrm{d}\frac{\partial u}{\partial t}(t, c(\alpha_3))(D_{\dot{c}(\alpha_3)}\dot{c}(\alpha_3))\right), \tag{B.5}$$

where $\alpha_3 \in (0, \sqrt{h})$.

Substituting (B.4) and (B.5) in (B.3), we arrive at

$$u(t+h, c(s)) = u + h\frac{\partial u}{\partial t} + s\mathrm{d}u(V) + \frac{s^2}{2}(D^2 u(V, V) + \mathrm{d}u(D_V V))$$

$$+ \frac{s^3}{6}(D^3 u(V, V, V) + 3D^2 u(V, D_V V) + \mathrm{d}u(D_V^2 V))$$

$$+ \frac{s^4}{24}\frac{\mathrm{d}^4}{\mathrm{d}s^4}u(t, c(\alpha_2))$$

$$+ hsD\frac{\partial u}{\partial t}(V) + h\frac{s^2}{2}\left(D^2\frac{\partial u}{\partial t}(t, c(\alpha_3))(\dot{c}(\alpha_3), \dot{c}(\alpha_3)) + \mathrm{d}\frac{\partial u}{\partial t}(t, c(\alpha_3))(D_{\dot{c}(\alpha_3)}\dot{c}(\alpha_3))\right)$$

$$+ \frac{h^2}{2} \frac{\partial^2 u}{\partial t^2}(t(\alpha_1), c(s)). \tag{B.6}$$

Taking expectation on both sides and putting $s = \sqrt{h}$, we get

$$\mathbb{E}[u(t+h, X_1^h)] = u + h\frac{\partial u}{\partial t} + \mathbb{E}\Bigg[\sqrt{h}\mathrm{d}u(V) + \frac{h}{2}(D^2 u(V,V) + \mathrm{d}u(D_V V))$$

$$+ \frac{h^{3/2}}{6}(D^3 u(V,V,V) + 3D^2 u(V, D_V V) + \mathrm{d}u(D_V^2 V)) + h^{3/2}D\frac{\partial u}{\partial t}(V)$$

$$+ \frac{h^2}{24}\frac{\mathrm{d}^4}{\mathrm{d}s^4}u(t, c(\alpha_2))$$

$$+ \frac{h^2}{2}\left(D^2\frac{\partial u}{\partial t}(t, c(\alpha_3))(\dot{c}(\alpha_3), \dot{c}(\alpha_3)) + \mathrm{d}\frac{\partial u}{\partial t}(t, c(\alpha_3))(D_{\dot{c}(\alpha_3)}\dot{c}(\alpha_3)) \right)$$

$$+ \frac{h^2}{2}\frac{\partial^2 u}{\partial t^2}(t(\alpha_1), c(s)) \Bigg]. \tag{B.7}$$

Denote i-th component of V as V^i, i.e. $V^i = \left(-\frac{h^{1/2}}{2}\nabla\phi + g^{-1/2}\xi \right)^i$. Using the properties of the random variables ξ^i

$$\mathbb{E}[\xi^i] = 0, \ \mathbb{E}[\xi^i\xi^j] = \delta_{ij}, \ \mathbb{E}[\xi^i\xi^j\xi^k] = 0, \ \mathbb{E}[(\xi^i)^2(\xi^j)^2] = 1 \text{ for } i \neq j, \tag{B.8}$$

we deduce that

$$\mathbb{E}(V^i V^j V^k) = -\frac{h^{3/2}}{8}(\nabla\phi)^i(\nabla\phi)^j(\nabla\phi)^k + \frac{3h}{4}(\nabla\phi)^i(\nabla\phi)^j\mathbb{E}\big((g^{-1/2}\xi)^k\big)$$

$$- \frac{3h^{1/2}}{2}(\nabla\phi)^i\mathbb{E}\big((g^{-1/2}\xi)^j(g^{-1/2}\xi)^k\big) + \mathbb{E}\big((g^{-1/2}\xi)^i(g^{-1/2}\xi)^j(g^{-1/2}\xi)^k\big)$$

$$= -\frac{h^{3/2}}{8}(\nabla\phi)^i(\nabla\phi)^j(\nabla\phi)^k - \frac{3h^{1/2}}{2}(\nabla\phi)^i\mathbb{E}\big((g^{1/2})^{jl}\xi_l(g^{1/2})^{km}\xi_m\big)$$

$$= -\frac{h^{3/2}}{8}(\nabla\phi)^i(\nabla\phi)^j(\nabla\phi)^k - \frac{3h^{1/2}}{2}(\nabla\phi)^i(g^{1/2})^{jl}(g^{1/2})^{km}\delta_{lm}$$

$$= -\frac{h^{3/2}}{8}(\nabla\phi)^i(\nabla\phi)^j(\nabla\phi)^k - \frac{3h^{1/2}}{2}(\nabla\phi)^i g^{jk}, \tag{B.9}$$

where $i, j, k = 1, \ldots, q$ and δ_{ij} is the Kronecker delta. Further, we have

$$\mathbb{E}(h^{1/2}Du(V)) = -\frac{h}{2}g(\nabla\phi, \nabla u), \tag{B.10}$$

$$\mathbb{E}(hD^2 u(V,V)) = h\mathbb{E}(\mathrm{Hess}^u(g^{-1/2}\xi, g^{-1/2}\xi)) + \frac{h^2}{4}D^2 u(\nabla\phi, \nabla\phi)$$

$$= h\mathbb{E}\big(\mathrm{Hess}_{ij}^u(g^{-1/2}\xi)^i(g^{-1/2}\xi)^j\big) + \frac{h^2}{4}D^2 u(\nabla\phi, \nabla\phi)$$

$$= h\mathbb{E}\big(\mathrm{Hess}_{ij}^u(g^{1/2})^{ik}\xi_k(g^{1/2})^{jl}\xi_l\big) + \frac{h^2}{4}D^2 u(\nabla\phi, \nabla\phi)$$

$$= h\mathrm{Hess}_{ij}^u(g^{1/2})^{ik}(g^{1/2})^{jl}\delta_{kl} + \frac{h^2}{4}D^2 u(\nabla\phi, \nabla\phi)$$

$$= h\text{Hess}_{ij}^u g^{ij} + \frac{h^2}{4} D^2 u(\nabla\phi, \nabla\phi)$$

$$= h\Delta_M u + \frac{h^2}{4} D^2 u(\nabla\phi, \nabla\phi), \tag{B.11}$$

$$\mathbb{E}\left(h^{3/2} D^3 u(V, V, V)\right) = h^{3/2} D^3 u(\partial_i, \partial_j, \partial_k) \mathbb{E}(V^i V^j V^k)$$

$$= -\frac{h^3}{8} D^3 u(\partial_i, \partial_j, \partial_k)(\nabla\phi)^i(\nabla\phi)^j(\nabla\phi)^k$$

$$- \frac{3h^2}{2} D^3 u(\partial_i, \partial_j, \partial_k)(\nabla\phi)^i g^{jk}, \tag{B.12}$$

$$\mathbb{E}\left(h^{3/2} D\frac{\partial u}{\partial t}(V)\right) = h^{3/2} D\frac{\partial u}{\partial t}(\partial_i)\mathbb{E}(V^i) = \frac{h^2}{2} D\frac{\partial u}{\partial t}(\partial_i)(\nabla\phi)^i. \tag{B.13}$$

Substituting (B.10)-(B.13) in (B.7), we get

$$\mathbb{E}(u(t+h, X_1^h)) = u + h\frac{\partial u}{\partial t} - \frac{h}{2}g(\nabla\phi, \nabla u) + \frac{h}{2}\Delta_M u$$

$$+ \frac{h^2}{8} D^2 u(\nabla\phi, \nabla\phi) - \frac{h^3}{48} D^3 u(\partial_i, \partial_j, \partial_k)(\nabla\phi)^i(\nabla\phi)^j(\nabla\phi)^k$$

$$- \frac{h^2}{4} D^3 u(\partial_i, \partial_j, \partial_k)(\nabla\phi)^i g^{jk} + \frac{h^2}{2} D\frac{\partial u}{\partial t}(\partial_i)(\nabla\phi)^i$$

$$+ \mathbb{E}\left(\frac{h^2}{24}\frac{\mathrm{d}^4}{\mathrm{d}s^4}u(t, c(\alpha_2))\right.$$

$$+ \frac{h^2}{2} D^2\frac{\partial u}{\partial t}(t, c(\alpha_3))(\dot c(\alpha_3), \dot c(\alpha_3)) + \frac{h^2}{2}\frac{\partial^2 u}{\partial t^2}(t(\alpha_1), c(s))$$

$$+ \mathrm{d}\frac{\partial u}{\partial t}(t, c(\alpha_3))(D_{\dot c(\alpha_3)}\dot c(\alpha_3))\right) + \frac{h}{2}\mathbb{E}(\mathrm{d}u(D_V V))$$

$$+ \frac{h^{3/2}}{6}\mathbb{E}(\mathrm{d}u(D_V^2 V)) + \frac{h^{3/2}}{2}\mathbb{E}(D^2 u(V, D_V V)). \tag{B.14}$$

We note that the remainder terms include a linear combination of derivatives of u which can be estimated using (2.14) together with boundedness of $\nabla\phi$ thanks to compactness of M and Assumptions 1 and 2. Due to Assumption 3 and again using Assumptions 1 and 2, the last three terms are also bounded by $Ce^{-\lambda(T-t)}h^2$, where C and λ are independent of h and T. Consequently, we arrive at (B.2). ∎

We prove the above lemma under the conditions that retraction satisfies Assumption 3. If we take $c(s)$ to be a geodesic curve then it is clear that $D_{\dot c(s)}\dot c(s)|_{s=0} = 0$ and $D_{\dot c(s)}D_{\dot c(s)}\dot c(s)|_{s=0} = 0$.

We now provide proof of Theorem 1.

B.1 Proof of Theorem 1

Note that using (2.12) and (2.15), we can write

$$|\mathbb{E}(\varphi(X_N^h)) - \mu_\phi(\varphi)| \leq |\mathbb{E}(\varphi(X_N^h)) - \mathbb{E}(\varphi(X(T)))| + |\mathbb{E}(\varphi(X(T))) - \mu_\phi(\varphi)|$$

$$= |\mathbb{E}(u(t_N, X_N^h)) - u(0, x_0)| + |\mathbb{E}(\varphi(X(T))) - \mu_\phi(\varphi)|$$

$$= \left| \mathbb{E}\left(\sum_{n=0}^{N-1} \mathbb{E}\left(u_{n+1} - u_n \mid X_n^h \right) \right) \right| + |\mathbb{E}(\varphi(X(T))) - \mu_\phi(\varphi)|.$$

Using Lemma 1 and (2.11), we get

$$|\mathbb{E}(\varphi(X_N^h)) - \mu_\phi(\varphi)| \le Ch^2 \sum_{n=0}^{N-1} e^{-\lambda(T-t_n)} + Ce^{-\lambda T} \le C(h + e^{-\lambda T}),$$

where $C > 0$ is a constant independent of h and T and it linearly depends on $\|\varphi\|_{C^{4,\epsilon}(M)}$ but otherwise C, λ are independent of φ. ∎

B.2 Proof of Theorem 2

Let us denote $u_n := u(X_n^h)$, $\nabla\phi_n := \nabla\phi(X_n^h)$, $g_n^{-1/2} = g^{-1/2}(X_n^h)$ and $V_n = -\frac{h^{1/2}}{2}\nabla\phi_n + g_n^{-1/2}\xi_{n+1}$, where $u(x)$ is the solution of (2.16). Define γ_{V_n} to be the geodesic with initial conditions

$$\gamma_{V_n}(0) = X_n^h, \tag{B.15}$$

$$\dot{\gamma}_{V_n}(0) = V_n := -\frac{h^{1/2}}{2}\nabla\phi_n + g_n^{-1/2}\xi_{n+1}. \tag{B.16}$$

Applying the covariant Taylor expansion (A.5) with $s = \sqrt{h}$ to u_{n+1} around X_n^h, we obtain

$$u_{n+1} - u_n = h^{1/2}Du_n(V_n) + \frac{h}{2}(D^2 u_n(V_n, V_n) + \mathrm{d}u_n(D_{V_n} V_n))$$

$$+ \frac{h^{3/2}}{6}(D^3 u_n(V_n, V_n, V_n) + 3D^2 u_n(V_n, D_{V_n} V_n) + \mathrm{d}u_n(D_{V_n}^2 V_n))$$

$$+ \frac{h^2}{24}\frac{\mathrm{d}^4}{\mathrm{d}s^4}u(\gamma_{V_n}(\alpha)), \tag{B.17}$$

where $\alpha \in (0, \sqrt{h})$. Taking expectation conditioned on X_n^h and using the fact that γ is a geodesic curve, we get

$$\mathbb{E}(u_{n+1} \mid X_n^h) = u_n - \frac{h}{2}g(\nabla\phi_n, \nabla u_n) + \frac{h}{2}\Delta_M u_n$$

$$+ \frac{h^2}{8}D^2 u_n(\nabla\phi_n, \nabla\phi_n) - \frac{h^3}{48}D^3 u_n(\partial_i, \partial_j, \partial_k)(\nabla\phi_n)^i(\nabla\phi_n)^j(\nabla\phi_n)^k$$

$$- \frac{h^2}{2}D^3 u_n(\partial_i, \partial_j, \partial_k)(\nabla\phi_n)^i g_n^{jk}$$

$$+ \frac{h^2}{24}\mathbb{E}[D^4 u(\gamma_{V_n}(\alpha))(\dot{\gamma}_{V_n}(\alpha), \dot{\gamma}_{V_n}(\alpha), \dot{\gamma}_{V_n}(\alpha), \dot{\gamma}_{V_n}(\alpha)) \mid X_n^h], \tag{B.18}$$

where ∂_i, $i = 1, \ldots, q$, denote basis vectors of $T_{X_n^h}M$ and $V_n \in T_{X_n^h}M$. The arguments employed to obtain (B.18) are based on (B.10)-(B.12) except that here we take expectation

conditional on X_n^h and u represents the solution of the Poisson PDE (2.16). Rearranging (B.18) and using the Poisson PDE (2.16), we arrive at

$$h\varphi(X_n^h) - h\mu_\phi(\varphi) = \mathbb{E}(u_{n+1}|X_n^h) - u_n - \frac{h^2}{8}D^2u_n(\nabla\phi_n, \nabla\phi_n)$$

$$+ \frac{h^3}{48}D^3u_n(\partial_i, \partial_j, \partial_k)(\nabla\phi_n)^i(\nabla\phi_n)^j(\nabla\phi_n)^k$$

$$+ \frac{h^2}{2}D^3u_n(\partial_i, \partial_j, \partial_k)(\nabla\phi_n)^i g_n^{jk}$$

$$- \frac{h^2}{24}\mathbb{E}[D^4u(\gamma_{V_n}(\alpha))(\dot\gamma_{V_n}(\alpha), \dot\gamma_{V_n}(\alpha), \dot\gamma_{V_n}(\alpha), \dot\gamma_{V_n}(\alpha)) \mid X_n^h]. \quad \text{(B.19)}$$

Taking expectation on both sides, summing over $n = 0, \ldots, N-1$, and applying (2.17), we ascertain

$$\left| h\mathbb{E}\sum_{n=0}^{N-1}\varphi(X_n^h) - hN\mu_\phi(\varphi) \right| \leq |\mathbb{E}(u_N - u_0)| + Ch^2N,$$

where $C > 0$ is independent of T and h and it linearly depends on $\|\varphi\|_{C^{2,\epsilon}(M)}$ but otherwise it is independent of φ. Dividing by T (note that $T = Nh$) gives us the result stated in (4.6):

$$|\mathbb{E}\tilde\mu_{\phi,N}(\varphi) - \mu_\phi(\varphi)| \leq C\left(h + \frac{1}{T}\right),$$

where we have used the fact that $|u_N - u_0|$ is bounded uniformly in x due to compactness of the manifold M and $u(x) \in C^{4,\epsilon}(M)$.

To show the bound (4.7), we start with (B.17)

$$u_{n+1} - u_n = h\mathcal{A}u_n + h^{1/2}Du_n(g_n^{-1/2}\xi_{n+1}) + \frac{h}{2}\left(D^2u_n(V_n, V_n) - \Delta_M u_n\right)$$

$$+ \frac{h^{3/2}}{6}D^3u_n(V_n, V_n, V_n) + \frac{h^2}{24}D^4u(\gamma_{V_n}(\alpha))(\dot\gamma_{V_n}(\alpha), \dot\gamma_{V_n}(\alpha), \dot\gamma_{V_n}(\alpha), \dot\gamma_{V_n}(\alpha)). \quad \text{(B.20)}$$

We rearrange (B.20), sum over $n = 0, \ldots, N-1$ and use (2.16) to arrive at

$$Nh(\tilde\mu_{\phi,N}(\varphi) - \mu_\phi(\varphi)) = u_N - u_0$$

$$+ \sum_{n=0}^{N-1}\left(-h^{1/2}Du_n(g_n^{-1/2}\xi_{n+1}) - \frac{h}{2}\left(D^2u_n(V_n, V_n) - \Delta_M u_n\right)\right.$$

$$- \frac{h^{3/2}}{6}D^3u_n(V_n, V_n, V_n)$$

$$\left. - \frac{h^2}{24}D^4u(\gamma_{V_n}(\alpha))(\dot\gamma_{V_n}(\alpha), \dot\gamma_{V_n}(\alpha), \dot\gamma_{V_n}(\alpha), \dot\gamma_{V_n}(\alpha))\right). \quad \text{(B.21)}$$

Squaring both sides and dividing by T^2, we obtain

$$(\tilde\mu_{\phi,N}(\varphi) - \mu_\phi(\varphi))^2 \leq \frac{C}{T^2}(u_N - u_0)^2 + \frac{Ch}{T^2}\left(\sum_{n=0}^{N-1}Du_n(g_n^{-1/2}\xi_{n+1})\right)^2$$

$$+ \frac{CNh^4}{T^2} \sum_{n=0}^{N-1} (D^2 u_n (\nabla \phi_n, \nabla \phi_n))^2$$

$$+ \frac{Ch^3}{T^2} \left(\sum_{n=0}^{N-1} D^2 u_n (g_n^{-1/2} \xi_{n+1}, \nabla \phi_n) \right)^2$$

$$+ \frac{Ch^3}{T^2} \left(\sum_{n=0}^{N-1} D^2 u_n (\nabla \phi_n, g_n^{-1/2} \xi_{n+1}) \right)^2$$

$$+ \frac{Ch^2}{T^2} \left(\sum_{n=0}^{N-1} D^2 u_n (g_n^{-1/2} \xi_{n+1}, g_n^{-1/2} \xi_{n+1}) - \Delta_M u_n \right)^2$$

$$+ \frac{CNh^6}{T^2} \sum_{n=0}^{N-1} (D^3 u_n (\nabla \phi_n, \nabla \phi_n, \nabla \phi_n))^2$$

$$+ \frac{CNh^5}{T^2} \sum_{n=0}^{N-1} (D^3 u_n (\partial_i, \partial_j, \partial_k)(\nabla \phi)^i (\nabla \phi)^j (g_n^{-1/2} \xi_{n+1})^k)^2$$

$$+ \frac{CNh^4}{T^2} \sum_{n=0}^{N-1} (D^3 u_n (\partial_i, \partial_j, \partial_k)(\nabla \phi)^i (g_n^{-1/2} \xi_{n+1})^j (g_n^{-1/2} \xi_{n+1})^k)^2$$

$$+ \frac{Ch^3}{T^2} \left(\sum_{n=0}^{N-1} D^3 u_n (\partial_i, \partial_j, \partial_k)(g_n^{-1/2} \xi_{n+1})^i (g_n^{-1/2} \xi_{n+1})^j (g_n^{-1/2} \xi_{n+1})^k \right)^2$$

$$+ \frac{CNh^4}{T^2} \sum_{n=0}^{N-1} \left(D^4 u(\gamma_{V_n}(\alpha))(\dot{\gamma}_{V_n}(\alpha), \dot{\gamma}_{V_n}(\alpha), \dot{\gamma}_{V_n}(\alpha), \dot{\gamma}_{V_n}(\alpha)) \right)^2. \tag{B.22}$$

Using (B.8), we get

$$\mathbb{E} \left(\sum_{n=0}^{N-1} Du_n (g_n^{-1/2} \xi_{n+1}) \right)^2 = \sum_{n=0}^{N-1} \mathbb{E}(Du_n (g_n^{-1/2} \xi_{n+1}))^2$$

$$+ 2 \sum_{k<n} \mathbb{E}(Du_k (g_k^{-1/2} \xi_{k+1}) Du_n (g_n^{-1/2} \xi_{n+1}))$$

$$= \sum_{n=0}^{N-1} \mathbb{E}(Du_n (g_n^{-1/2} \xi_{n+1}))^2, \tag{B.23}$$

and

$$\mathbb{E} \left(\sum_{n=0}^{N-1} D^2 u_n (g_n^{-1/2} \xi_{n+1}, \nabla \phi_n) \right)^2 = \sum_{n=0}^{N-1} \mathbb{E}(D^2 u_n (g_n^{-1/2} \xi_{n+1}, \nabla \phi_n))^2. \tag{B.24}$$

Analogously,

$$\mathbb{E} \left(\sum_{n=0}^{N-1} D^3 u_n (\partial_i, \partial_j, \partial_k)(g_n^{-1/2} \xi_{n+1})^i (g_n^{-1/2} \xi_{n+1})^j (g_n^{-1/2} \xi_{n+1})^k \right)^2$$

$$= \sum_{n=0}^{N-1} \mathbb{E}(D^3 u_n(\partial_i, \partial_j, \partial_k)(g_n^{-1/2}\xi_{n+1})^i (g_n^{-1/2}\xi_{n+1})^j (g_n^{-1/2}\xi_{n+1})^k)^2 \qquad \text{(B.25)}$$

Noting that $\mathbb{E}[D^2 u_n(g_n^{-1/2}\xi_{n+1}, g_n^{-1/2}\xi_{n+1}) - \Delta_M u_n \mid X_n^h] = 0$, we get

$$\mathbb{E}\left(\sum_{n=0}^{N-1} D^2 u_n(g_n^{-1/2}\xi_{n+1}, g_n^{-1/2}\xi_{n+1}) - \Delta_M u_n \right)^2$$
$$= \sum_{n=0}^{N-1} \mathbb{E}(D^2 u_n(g_n^{-1/2}\xi_{n+1}, g_n^{-1/2}\xi_{n+1}) - \Delta_M u_n)^2. \qquad \text{(B.26)}$$

Taking expectation on both sides of (B.22), using (B.23)-(B.26) and applying (2.17), we obtain

$$\mathbb{E}(\tilde{\mu}_{\phi,N}(\varphi) - \mu_\phi(\varphi))^2 \leq C\left(h^2 + \frac{1}{T} \right),$$

where $C > 0$ is a constant independent of h and T and it linearly depends on $\|\varphi\|_{C^{2,\epsilon}(M)}$ but otherwise it is independent of φ. ∎

B.3 Proof of Theorem 3

Consider $\varphi \in \mathcal{H}$ (see (4.8)). Since μ_ϕ^h is stationary measure of Markov chain (see Section 3), we have

$$\int_M \varphi(x)\mu_\phi^h(\mathrm{d}x) = \int_M \mathbb{E}_x\varphi(X_k^h)\mu_\phi^h(\mathrm{d}x),$$

where $\mathbb{E}_x\varphi(X_k^h)$ implies that Markov chain starts at x, i.e. $X_0 = x$. Therefore, we can write

$$\int_M \varphi(x)\mu_\phi^h(\mathrm{d}x) = \int_M \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}_x\varphi(X_k^h)\mu_\phi^h(\mathrm{d}x).$$

Using Theorem 2, we get

$$\left| \int_M \varphi(x)\big(\mu_\phi^h(\mathrm{d}x) - \mu(\mathrm{d}x)\big) \right| = \left| \int_M \left(\frac{1}{N} \sum_{k=1}^{N} \mathbb{E}_x\varphi(X_k^h) - \mu_\phi(\varphi) \right)\mu_\phi^h(\mathrm{d}x) \right| \leq K\left(h + \frac{1}{Nh} \right),$$

where $K > 0$ is independent of h, N and φ from $\mathcal{H}$ (the latter thanks to the constant C in Theorem 2 being dependent on φ only via linear dependence on $\|\varphi\|_{C^{2,\epsilon}(M)}$ which is bounded for the class $\mathcal{H}$ of functions φ). Letting $N \to \infty$, we get the desired result. ∎

References

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds.* Princeton University Press, 2008.

Au. Alfonsi, D. Krief, and P. Tankov. Long-time large deviations for the multiasset Wishart stochastic volatility model and option pricing. *SIAM J. Finan. Math.*, 10(4):942–976, 2019.

J. Armstrong and T. King. Curved schemes for stochastic differential equations on, or near, manifolds. *Proceedings Royal Society A*, 478(2262):20210785, 2022.

T. Aubin. *Some Nonlinear Problems in Riemannian Geometry*. Springer, Berlin, 1998.

D. Bakry. Un Critère de non-explosion pour certaines diffusions sur une variété Riemannienne complète. *Comptes rendus de l'Académie des sciences. Série 1, Mathématique*, 303: 23–26, 1986.

D. Bakry and M. Émery. Diffusions hypercontractives. In J. Azéma and M. Yor, editors, *Séminaire de Probabilités XIX 1983/84*, pages 177–206. Springer, 1985.

V. Bally and D. Talay. The Euler scheme for stochastic differential equations: error analysis with Malliavin calculus. *Math. Computers Simul.*, 38(1):35–41, 1995.

A. Barp, L. Da Costa, G. Franca, K. Friston, M. Girolami, M.I. Jordan, and G.A. Pavliotis. Geometric methods for sampling, optimisation, inference and adaptive agents. In *Handbook of Statistics*, volume 46, pages 21–78. World Sci., 2022.

K. Bharath, A. Lewis, and M. V. Tretyakov. Fundamental theorem for mean-square approximation of SDEs on manifolds. *In preparation*, 2025.

R. L. Bishop and R. J. Crittenden. *Geometry of Manifolds*. Academic Press, 2011.

N. Bou-Rabee and M. Hairer. Non-asymptotic mixing of the MALA algorithm. *IMA J. Numer. Anal.*, 33(1):80–110, 2013.

M.-F. Bru. Wishart processes. *J. Theor. Probab.*, 4:725–751, 1991.

J. Cheeger and D.G. Ebin. *Comparison Theorems in Riemannian Geometry*. North-Holland Publishing Company, 1975.

X. Cheng, J. Zhang, and S. Sra. Efficient sampling on Riemannian manifolds via Langevin MCMC. *Advances in Neural Information Processing Systems*, 35:5995–6006, 2022.

A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. Royal Stat. Society B*, 79(3):651–676, 2017.

G. Da Prato and J. Zabczyk. *Ergodicity for Infinite Dimensional Systems*. Cambr. Univ. Press, 1996.

R. L. Davidchack, R. Handel, and M. V. Tretyakov. Langevin thermostat for rigid body dynamics. *J. Chem. Phys.*, 130(23):234101, 2009.

R. L. Davidchack, T. E. Ouldridge, and M. V. Tretyakov. New Langevin and gradient thermostats for rigid body dynamics. *J. Chem. Phys.*, 142(14):144114, 2015.

T. D. Downs. Orientation statistics. *Biometrika*, 59(3):665–676, 1972.

A. Durmus and É. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551 – 1587, 2017.

K. D. Elworthy. *Stochastic Differential Equations on Manifolds.* London Mathematical Society Lecture Note Series. Cambridge University Press, 1982.

B. Everitt and C. Maclachlan. Constructing hyperbolic manifolds. *Computational and Geometric Aspects of Modern Algebra. Michael Atkinson et al.(Editors), London Maths. Soc. Lect. Notes*, 275:78–86, 2000.

N.I. Fisher, T. Lewis, and B.J.J. Embleton. *Statistical Analysis of Spherical Data.* Cambridge University Press, 1993.

J. Da Fonseca, M. Grasselli, and C. Tebaldi. A multifactor volatility Heston model. *Quant. Finance*, 8(6):591–604, 2008.

M. I. Freidlin. *Functional Integration and Partial Differential Equations.* Princeton Univ. Press, 1985.

S. Gallot, D. Hulin, and J. Lafontaine. *Riemannian Geometry*, volume 2. Springer, 1990.

K. Gatmiry and S. S. Vempala. Convergence of the Riemannian Langevin algorithm. *arXiv:2204.10818*, 2022.

M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. Royal Stat. Society B*, 73(2):123–214, 2011.

R. E. Greene and H.-H. Wu. *Function theory on Manifolds Which Possess a Pole.* Springer, 2006.

A. Grorud and D. Talay. Approximation of Lyapunov exponents of nonlinear stochastic differential equations. *SIAM J. Appl. Math.*, 56(2):627–650, 1996.

E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations. I. Nonstiff Problems.* Springer, Berlin, 1993.

E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration: Structure Presesrving Algorithms for Ordinary Differential Equations.* Springer, Berlin, 2002.

E. P. Hsu. *Stochastic Analysis on Manifolds.* Amer. Math. Soc., 2002.

M. Hutzenthaler and A. Jentzen. *Numerical approximations of stochastic differential equations with non-globally Lipschitz continuous coefficients*, volume 236 of *Mem. Amer. Math. Soc.* AMS, Providence, 2015.

M. Hutzenthaler, A. Jentzen, and P. E. Kloeden. Strong and weak divergence in finite time of Euler's method for stochastic differential equations with non-globally Lipschitz continuous coefficients. *Proc. R. Soc. A*, 467:1563–1576, 2011.

K. Ichihara. Curvature, geodesics and the Brownian motion on a Riemannian manifold i—recurrence properties. *Nagoya Math. J.*, 87:101–114, 1982a.

K. Ichihara. Curvature, geodesics and the Brownian motion on a Riemannian manifold ii—explosion properties. *Nagoya Math. J.*, 87:115–125, 1982b.

N. Ikeda and S. Watanabe. *Stochastic Differential Equations and Diffusion Processes.* Elsevier, 2014.

J. Jost. *Riemannian Geometry and Geometric Analysis.* Universitext. Springer, 2008.

P. E. Jupp and K. V. Mardia. Maximum likelihood estimators for the matrix von Mises-Fisher and Bingham distributions. *Ann. Stat.*, 7(3):599–606, 1979.

W. S. Kendall. Brownian motion on a surface of negative curvature. *Séminaire de probabilités de Strasbourg*, 18:70–76, 1984.

R. Khasminskii. *Stochastic Stability of Differential Equations.* Springer, 2012.

S. Kobayashi and K. Nomizu. *Foundations of Differential Geometry, Vol II.* John Wiley & Sons, 1969.

G. Kurz and U.D. Hanebeck. Stochastic sampling of the hyperspherical von Mises–Fisher distribution without rejection methods. In *2015 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, pages 1–6. IEEE, 2015.

O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Ural'tseva. *Linear and quasi-linear equations of parabolic type*, volume 23 of *Trans. Math. Monog.* Amer. Math. Soc., Providence, RI, 1968.

A. Laurent and G. Vilmart. Order conditions for sampling the invariant measure of ergodic stochastic differential equations on manifolds. *Found. Comput. Math.*, 22:649–695, 2022.

H. Le, A. Lewis, K. Bharath, and C. Fallaize. A diffusion approach to Stein's method on Riemannian manifolds. *Bernoulli*, 30(2):1079–1104, 2024.

B. Leimkuhler and C. Matthews. *Molecular Dynamics with Deterministic and Stochastic Numerical Methods.* Springer, Berlin, 2015.

B. Leimkuhler, A. Sharma, and M. V. Tretyakov. Simplest random walk for approximating Robin boundary value problems and ergodic limits of reflected diffusions. *Ann. Appl. Probab.*, 33(3):1904 – 1960, 2023.

T. Leliévre, M. Rousset, and G. Stoltz. Langevin dynamics with constraints and computation of free energy differences. *Math. Comp.*, 81(280):2071–2125, 2012.

A. Lewis. *Contributions to Stein's Method on Riemannian Manifolds.* PhD thesis, University of Nottingham, 2023.

M. B. Li and M. A. Erdogdu. Riemannian Langevin algorithm for solving semidefinite programs. *Bernoulli*, 29:3093 – 3113, 2023.

X.-M. Li. *Stochastic Flows on Non-compact Manifolds*. PhD thesis, University of Warwick, 1992.

A. Lunardi. *Analytic Semigroups and Optimal Regularity in Parabolic Problems*. Springer, 1995.

S. J. A. Malham and A. Wiese. Stochastic Lie group integrators. *SIAM J. Sci. Comp.*, 30 (2):597–617, 2008.

O. Mangoubi and A. Smith. Rapid mixing of geodesic walks on manifolds with positive curvature. *Ann. Appl. Probab.*, 28(4):2501–2543, 2018.

J. C. Mattingly, A. M. Stuart, and D. J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stoch. Proc. Appl.*, 101:185–232, 2002.

J. C. Mattingly, A. Stuart, and M. V. Tretyakov. Convergence of numerical time-averaging and stationary measures via Poisson equations. *SIAM J. Numer. Anal.*, 48(2):552–577, 2010.

J. H. Mentink, M. V. Tretyakov, A. Fasolino, M. I. Katsnelson, and Th. Rasing. Stable and fast semi-implicit integration of the stochastic Landau-Lifshitz equation. *J. Phys.: Condens. Matter*, 22(17):176001, 2010.

A. Mijatović, V. Mramor, and G. U. Bravo. A note on the exact simulation of spherical Brownian motion. *Stat. & Prob. Lett.*, 165:108836, 2020.

G. N. Milstein. A method with second order accuracy for the integration of stochastic differential equations. *Theor. Prob. Appl.*, 23:414–419, 1978.

G. N. Milstein. Weak approximation of solutions of systems of stochastic differential equations. *Theor. Prob. Appl.*, 30:706–721, 1985.

G. N. Milstein and M. V. Tretyakov. Simulation of a space-time bounded diffusion. *Ann. Appl. Probab.*, 9(3):732–779, 1999.

G. N. Milstein and M. V. Tretyakov. Numerical integration of stochastic differential equations with nonglobally Lipschitz coefficients. *SIAM J. Numer. Anal.*, 43(3):1139–1154, 2005.

G. N. Milstein and M. V. Tretyakov. Computing ergodic limits for Langevin equations. *Physica D*, 229(1):81–95, 2007.

G. N. Milstein and M. V. Tretyakov. *Stochastic Numerics for Mathematical Physics*. Springer, 2nd edition, 2021.

C. Miranda. *Partial Differential Equations of Elliptic Type*. Springer, 1969.

M. Moakher and M. Zéraï. The Riemannian geometry of the space of positive-definite matrices and its application to the regularization of positive-definite matrix-valued data. *J. Mathem. Imaging and Vision*, 40:171–187, 2011.

L. I. Nicolaescu. *Lectures on the Geometry of Manifolds*. World Scientific, 2007.

X. Pennec, P. Fillard, and N. Ayache. A Riemannian framework for tensor computing. *Inter. J. Computer Vision*, 66:41–66, 2006.

G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2:341–363, 1996.

S. Said, L. Bombrun, Y. Berthoumieu, and J. H. Manton. Riemannian Gaussian distributions on the space of symmetric positive definite matrices. *IEEE Transactions on Information Theory*, 63:2153–2170, 2017.

S. Schwarz, M. Herrmann, A. Sturm, and M. Wardetzky. Efficient random walks on Riemannian manifolds. *Found. Comput. Mathem.*, pages 1–17, 2023.

U. Sharma and W. Zhang. Nonreversible sampling schemes on submanifolds. *SIAM J. Numer. Anal.*, 59(6):2989–3031, 2021.

D. Talay. Discrétisation d'une équation différentielle stochastique et calcul approché d'espérances de fonctionnelles de la solution. *Math. Model. Numer. Anal. (ESAIM)*, 20(1):141–179, 1986.

D. Talay. Second-order discretization schemes of stochastic differential systems for the computation of the invariant law. *Stoch. Stoch. Reports*, 29:13–36, 1990.

D. Talay and L. Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stoch. Anal. Appl.*, 8:483–509, 1990.

Y. W. Teh, A. H. Thiery, and S. J. Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *J. Machine Learning Research*, 17(7):1–33, 2016.

M. V. Tretyakov. Sampling from mixture distributions based on regime-switching diffusions. *SIAM J. Sci. Comp.*, to appear, 2025.

M. V. Tretyakov and Z. Zhang. A fundamental mean-square convergence theorem for SDEs with locally Lipschitz coefficients and its applications. *SIAM J. Numer. Anal.*, 51:3135–3162, 2013.

F.-Y. Wang. Log-Sobolev inequalities: different roles of Ric and Hess. *Ann. Probab.*, 37(4):1587–1604, 2009.

F.-Y. Wang. *Analysis for Diffusion Processes on Riemannian Manifolds*. World Scientific, 2013.

G. S. Watson. Equatorial distributions on a sphere. *Biometrika*, 52(1/2):193–201, 1965.

A. T. A. Wood. Simulation of the von Mises Fisher distribution. *Commun. Statistics-Simulation Comput.*, 23(1):157–164, 1994.