Scandinavian Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics, Vol. 40: 825–845, 2013
doi: 10.1111/sjos.12036

Geodesic Monte Carlo on Embedded Manifolds

SIMON BYRNE and MARK GIROLAMI

Department of Statistical Science, University College London

ABSTRACT. Markov chain Monte Carlo methods explicitly defined on the manifold of probability distributions have recently been established. These methods are constructed from diffusions across the manifold and the solution of the equations describing geodesic flows in the Hamilton–Jacobi representation. This paper takes the differential geometric basis of Markov chain Monte Carlo further by considering methods to simulate from probability distributions that themselves are defined on a manifold, with common examples being classes of distributions describing directional statistics. Proposal mechanisms are developed based on the geodesic flows over the manifolds of support for the distributions, and illustrative examples are provided for the hypersphere and Stiefel manifold of orthonormal matrices.

Key words: directional statistics, geodesic, Hamiltonian Monte Carlo, Riemannian manifold, Stiefel manifold

1. Introduction

Markov chain Monte Carlo (MCMC) methods that originated in the physics literature have caused a revolution in statistical methodology over the last 20 years by providing the means, now in an almost routine manner, to perform Bayesian inference over arbitrary non-conjugate prior and posterior pairs of distributions (Gilks *et al.*, 1996).

A specific class of MCMC methods, originally known as hybrid Monte Carlo (HMC), was developed to more efficiently simulate quantum chromodynamic systems (Duane *et al.*, 1987). HMC goes beyond the random walk Metropolis or Gibbs sampling schemes and overcomes many of their shortcomings. In particular, HMC methods are capable of proposing bold long distance moves in the state space that will retain a very high acceptance probability and thus improve the rate of convergence to the invariant measure of the chain and reduce the autocorrelation of samples drawn from the stationary distribution of the chain. The HMC proposal mechanism is based on simulating Hamiltonian dynamics defined by the target distribution (see Neal (2011) for a comprehensive tutorial). For this reason, HMC is now routinely referred to as Hamiltonian Monte Carlo. Despite the relative strengths and attractive properties of HMC, it has largely been bypassed in the literature devoted to MCMC and Bayesian statistical methodology with very few serious applications of the methodology being published.

More recently, Girolami & Calderhead (2011) defined a Hamiltonian scheme that is able to incorporate geometric structure in the form a Riemannian metric. The Riemannian manifold Hamiltonian Monte Carlo (RMHMC) methodology makes proposals implicitly via Hamiltonian dynamics on the manifold defined by the Fisher–Rao metric tensor and the corresponding Levi-Civita connection. The paper has raised an awareness of the differential geometric foundations of MCMC schemes such as HMC and has already seen a number of methodological and algorithmic developments as well as some impressive and challenging applications exploiting these geometric MCMC methods (Konukoglu *et al.*, 2011; Martin *et al.*, 2012; Raue *et al.*, 2012; Vanlier *et al.*, 2012).

In contrast to Girolami & Calderhead (2011), in this particular paper, we show how Hamiltonian Monte Carlo methods may be designed for and applied to distributions defined on manifolds embedded in Euclidean space, by exploiting the existence of explicit forms for geodesics. This can provide a significant boost in speed, by avoiding the need to solve large linear systems as well as complications arising because of the lack of a single global coordinate system.

By way of specific illustration, we consider two such manifolds: the unit hypersphere, corresponding to the set of unit vectors in $\mathbb{R}^d$, and its extension to Stiefel manifolds, the set of p-tuples of orthogonal unit vectors in $\mathbb{R}^d$. Such manifolds occur in many statistical applications: distributions on circles and spheres, such as the von Mises distribution, are common in problems dealing with directional data (Mardia and Jupp, 2000). Orthonormal bases arise in dimension reduction methods such as factor analysis (Jolliffe, 1986) and can be used to construct distributions on matrices via eigendecompositions.

The problem of sampling from such distributions has not received much attention. Most methods in wide use, such as those used in directional statistics for sampling from spheres, have been developed for the specific problem at hand, often based on rejection sampling techniques tuned to a specific family. For the various multivariate extensions of these distributions, these techniques are usually embedded in a Gibbs sampling scheme.

There are relatively few works on the general problem of sampling from manifolds. The recent paper by Diaconis *et al.* (2013) provides a readable introduction to the concepts of geometric measure theory, and practical issues when sampling from manifolds, with the motivation of computing certain sampling distributions for hypothesis testing. Brubaker *et al.* (2012), somewhat similar to our approach, develop an HMC algorithm using the iterative algorithm for approximating the Hamiltonian paths.

In the next section, we provide a brief overview of the necessary concepts from differential geometry and geometric measure theory, such as geodesics and Hausdorff measures. In section 3, we construct a Hamiltonian integrator that utilizes the explicit form of the geodesics and incorporate this into a general HMC algorithm. Section 4 gives examples of various manifolds for which the geodesic equations are known, and section 5 provides some illustrative applications.

2. Manifolds, geodesics and measures

2.1. Manifolds and embeddings

In this section, we introduce the necessary terminology from differential geometry and information geometry. A more rigorous treatment can be found in reference books such as do Carmo (1976, 1992) and Amari & Nagaoka (2000).

An m-dimensional manifold $\mathcal{M}$ is a set that locally acts like $\mathbb{R}^m$: that is, for each point $x \in \mathcal{M}$, there is a bijective mapping q, called a *coordinate system*, from an open set around x to an open set in $\mathbb{R}^m$. Our particular focus is on manifolds that are *embedded* in some higher-dimensional Euclidean space $\mathbb{R}^n$, (i.e. they are submanifolds of $\mathbb{R}^n$). Note that $\mathbb{R}^d$ is itself a d-dimensional manifold, which we refer to as the *Euclidean manifold*.

Example 2.1. A simple example of an embedded manifold is the *hypersphere* or $(d-1)$-*sphere*:

$$\mathbb{S}^{d-1} = \left\{ x \in \mathbb{R}^d : \|x\| = 1 \right\}.$$

This is a $(d-1)$-dimensional manifold, as there exists an angular coordinate system $\phi \in (0, 2\pi) \times (0, \pi)^{d-2}$ where

$$
\begin{aligned}
x_1 &= \sin\phi_1 \ldots \sin\phi_{n-2} \sin\phi_{n-1},\\
x_2 &= \sin\phi_1 \ldots \sin\phi_{n-2} \cos\phi_{n-1},\\
x_3 &= \sin\phi_1 \ldots \cos\phi_{n-2},\\
&\vdots\\
x_{n-1} &= \sin\phi_1 \cos\phi_2,\\
x_n &= \cos\phi_1.
\end{aligned}
$$

Note that this coordinate system excludes some points of $\mathbb{S}^{d-1}$: such as $\delta_d = (0, \ldots, 0, 1)$. As a result, it is not a *global* coordinate system (in fact, no global coordinate system for $\mathbb{S}^{d-1}$ exists); nevertheless, it is possible to cover all of $\mathbb{S}^{d-1}$ by utilizing multiple coordinate systems known as an *atlas*.

A *tangent* at a point $x \in \mathcal{M}$ is a vector v that lies 'flat' on the manifold. More precisely, it can be defined as an equivalence class of the set of functions $\{\gamma : [a, b] \to \mathcal{M} : \gamma(t_0) = x\}$ that have the same 'time derivative' $\frac{\mathrm{d}}{\mathrm{d}t} q(\gamma(t))|_{t=t_0}$ in some coordinate system q. For an embedded manifold, however, a tangent can be represented simply as a vector $v \in \mathbb{R}^n$ such that

$$v = \dot{\gamma}(t_0) = \frac{\mathrm{d}}{\mathrm{d}t}\gamma(t)|_{t=t_0}.$$

The *tangent space* is the set T_x of such vectors and form a subspace of $\mathbb{R}^n$: this is equal to the span of the set of partial derivatives $\partial x_i / \partial q_j$ of some coordinate system q.

Example 2.2. A function on the sphere $\gamma : [a, b] \to \mathbb{S}^{d-1}$ must satisfy the constraint $\sum_i [\gamma_i(t)]^2 = 1$. By taking the time derivative of both sides, we find that

$$\frac{\mathrm{d}}{\mathrm{d}t} \sum_{i=1}^{d} [\gamma_i(t)]^2 = 2 \sum_{i=1}^{d} \gamma_i(t)\dot{\gamma}_i(t) = 0.$$

Therefore, the tangent space at $x \in \mathbb{S}^{d-1}$ is the $(d-1)$-dimensional subspace of vectors orthogonal to x:

$$T_x = \{v \in \mathbb{R}^d : x^\top v = 0\}.$$

A *Riemannian manifold* incorporates a notion of distance, such that for a point $q \in \mathcal{M}$, there exists a positive-definite matrix G, called the metric tensor, that forms an inner product between tangents u and v

$$\langle u, v \rangle_G = u^\top G(q) v.$$

Information geometry is the application of differential geometry to families of probability distributions. Such a family $\{p(\cdot \mid \theta) : \theta \in \Theta\}$ can be viewed as a Riemannian manifold, using the *Fisher–Rao metric tensor*

$$G_{ij} = -E_{X|\theta}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(X \mid \theta)\right].$$

Example 2.3. The family of d-dimensional multinomial distributions

$$p(z \mid \theta) = \theta_1^{z_1} \cdot \ldots \cdot \theta_d^{z_d}, \quad z = \delta_1, \ldots, \delta_d,$$

where δ_i is the ith coordinate vector, is parametrized by the unit $(d-1)$-simplex,

$$\Delta^{d-1} = \left\{ \theta \in \mathbb{R}^d : \theta_i \geq 0, \sum_j \theta_j = 1 \right\}.$$

This is a $(d-1)$-dimensional manifold embedded in $\mathbb{R}^d$ and can be parametrized in $(d-1)$ dimensions by dropping the last element of θ, the set of which we will denote by $\Delta^{d-1}_{(-d)}$.

The Fisher–Rao metric tensor in $\Delta^{d-1}_{(-d)}$ is then easily shown to be

$$G_{ij} \begin{cases} \frac{1}{\theta_i} + \frac{1}{1-\sum_{k=1}^{d-1}\theta_k} & \text{if } i = j, \\ \frac{1}{1-\sum_{k=1}^{d-1}\theta_k} & \text{otherwise.} \end{cases}$$

A smooth mapping from a Riemannian manifold to $\mathbb{R}^n$ is an *isometric embedding* if the Riemannian inner product is equivalent to the usual Euclidean inner product. That is,

$$s^\top G(q) t = u^\top v, \qquad \text{where } u_i = \sum_j \frac{\partial x_i}{\partial q_j} s_i,\ v_i = \sum_j \frac{\partial x_i}{\partial q_j} t_i,$$

or equivalently,

$$G_{ij} = \sum_{l=1}^{d} \frac{\partial x_l}{\partial q_i} \frac{\partial x_l}{\partial q_j}. \tag{1}$$

The existence of such embeddings is determined by the celebrated Nash (1956) embedding theorem; however, it does not give any guide as how to construct them. Nevertheless, there are some such embeddings we can identify.

Example 2.4. There is a bijective mapping from the simplex Δ^{d-1} to the positive orthant of the sphere $\mathbb{S}^{d-1}$ by taking the element-wise square root $x_i = \sqrt{\theta_i}$ (Fig. 1). If we consider it as a mapping from $\Delta^{d-1}_{(-d)}$, then the partial derivatives are of the form

$$\frac{\partial x_l}{\partial \theta_i} = \begin{cases} \frac{1}{2}\theta_i^{-1/2} & \text{if } i = l < d, \\ 0 & \text{if } i \neq l < d, \\ \frac{1}{2}\left(1 - \sum_{k=1}^{d-1}\theta_k\right)^{-1/2} & \text{if } i = l = d. \end{cases}$$

Note that by (1), this is an isometric embedding (up to proportionality) of the Fisher–Rao metric from example 2.3.

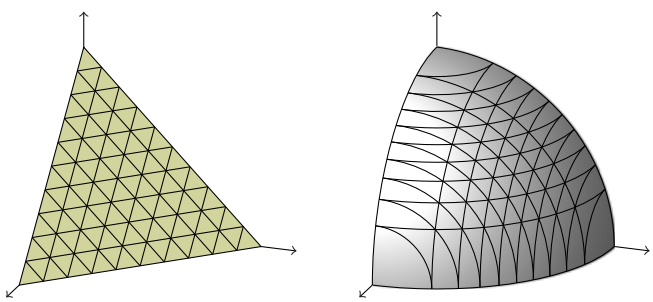

Fig. 1. Unit 2-simplex Δ^2 and the positive orthant of the two-sphere $\mathbb{S}^2$. The lines on the simplex are equidistant: the transformation to the sphere stretches these apart near the boundary.

2.2. Geodesics

The *affine connection* of a manifold determines the relationship between tangent spaces of different points on a manifold: interestingly, this depends on the path $\gamma : [a, b] \to \mathcal{M}$ used to connect the two points, and for a vector field $v(t) \in T_{\gamma(t)}$ along the path, we can measure the change by the *covariant derivative*.

Of course, the time derivative $\dot{\gamma}(t) = \frac{\mathrm{d}\gamma(t)}{\mathrm{d}t}$ is itself such a vector field: when this follows the affine connection, the covariant derivative is 0; in which case, γ is known as a *geodesic*.

This property can be expressed by the geodesic equation

$$\ddot{\gamma}_i(t) + \sum_{j,k} \Gamma^i_{jk}(\gamma(t))\, \dot{\gamma}_j(t) \dot{\gamma}_k(t) = 0, \tag{2}$$

where $\Gamma^i_{jk}(x)$ are known as the connection coefficients or Christoffel symbols. A Riemannian manifold induces a natural affine connection known as the *Levi-Civita connection*.

In the Euclidean manifold $\mathbb{R}^n$, the Christoffel symbols Γ^i_{jk} are zero, and so the geodesic (2) reduces to $\ddot{\gamma}(t) = 0$. Hence, the geodesics are the set of straight lines $\gamma(t) = at + b$.

In a Riemannian manifold, the geodesics are the locally extremal paths (maxima or minima in terms of calculus of variations) of the integrated path length

$$\int_a^b \|\dot{\gamma}(t)\|_G \, \mathrm{d}t, \quad \text{where } \|v\|_G^2 = v^\top G v.$$

Moreover, the geodesics have *constant speed*, in that $\|\dot{\gamma}(t)\|_G$ is constant over t. As the geodesics can be determined by the metric, they are consequently preserved under any metric-preserving transformation, such as an isometric embedding.

Example 2.5. A standard result in differential geometry is that the geodesics of the n-sphere are rotations about the origin, known as *great circles* (Fig. 2):

$$x(t) = x(0)\cos(\alpha t) + \frac{v(0)}{\alpha}\sin(\alpha t),$$

where $x(0) \in \mathbb{S}^n$ is the initial position, $v(0)$ is the initial velocity in the tangent space (i.e. such that $x(0)^\top v(0) = 0$) and $\alpha = \|v(0)\|$ is the constant angular velocity.

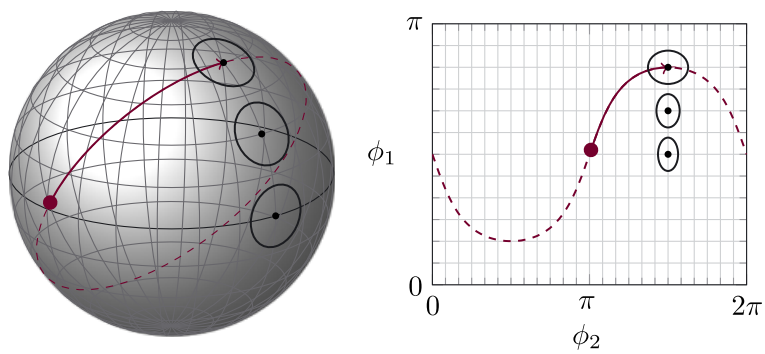

Fig. 2. A geodesic (⟶) and great circle (- - -) on the sphere $\mathbb{S}^2$ and its path in the spherical polar coordinate system $x = (\sin\phi_1 \sin\phi_2, \sin\phi_1 \cos\phi_2, \cos\phi_1)$. The ellipses correspond to equi-length tangents from each marked point.

For any geodesic $\gamma : [a, b] \to \mathcal{M}$, the *geodesic flow* describes the path of the geodesic and its tangent $(\gamma(t), \dot{\gamma}(t))$. Moreover, it is unique to the initial conditions $(x, v) = (\gamma(a), \dot{\gamma}(a))$, so we can describe any geodesic flow from its starting position x and velocity v: this is also known as the *exponential map*. If all such pairs (x, v) describe geodesics, then the manifold is said to be *geodesically complete*, which is true of the manifolds we consider in this paper.

2.3. The Hausdorff measure and distributions on manifolds

As our motivation is to sample from distributions defined on manifolds, we introduce some basic concepts of geometric measure theory that will be useful for this purpose. Geometric measure theory is a large and active topic and is covered in detail in references such as Federer (1969) and Morgan (2009). However, for a more accessible overview with a statistical flavour, we suggest the recent introduction given by Diaconis *et al.* (2013).

Our key requirement is a reference measure from which we can specify probability density functions, similar to the role played by the Lebesgue measure for distributions on Euclidean space. For this, we use the *Hausdorff measure*, one of the fundamental concepts in geometric measure theory. This can be defined rigorously in terms of a limit of coverings of the manifold (see the aforementioned references); however, for a manifold embedded in $\mathbb{R}^n$, it can be heuristically interpreted as the surface area of the manifold.

The relationship between $\mathcal{H}^m$, the m-dimensional Hausdorff measure and λ^m, the Lebesgue measure on $\mathbb{R}^m$, is given by the *area formula* (Federer, 1969, theorem 3.2.5). If we parametrize the manifold by a Libschitz function $f : \mathbb{R}^m \to \mathbb{R}^n$, then for any $\mathcal{H}^m$-measurable function $g : \mathbb{R}^n \to \mathbb{R}$,

$$\int_A g(f(u))\, J_m f(u)\, \lambda^m(\mathrm{d}u) = \int_{R^n} g(x)\, |\{u \in A : f(u) = x\}|\, \mathcal{H}^m(\mathrm{d}x).$$

Here, $J_m f(x)$ is the m-dimensional Jacobian of f: this can be defined as a norm on the matrix of partial derivatives $Df(x)$ (Federer, 1969, section 3.2.1), and if rank $Df(x) = m$, then $[J_m f(x)]^2$ is equal to the sum of squares of the determinants of all $m \times m$ submatrices of $Df(x)$.

Example 2.6. The square root mapping in example 2.4 from $\Delta_{(-d)}^{d-1}$ to $\mathbb{S}^{d-1}$ has $(d-1)$-dimensional Jacobian

$$\frac{1}{2^{d-1}} \prod_{i=1}^{d} \theta_i^{-1/2}.$$

The Dirichlet distribution is a distribution on the simplex, with density

$$\frac{1}{B(\alpha)} \prod_{i=1}^{d} \theta_i^{\alpha_i - 1}$$

with respect to the Lebesgue measure on $\Delta_{(-d)}^{d-1}$. Therefore, the corresponding density with respect to the Hausdorff measure on $\mathbb{S}^{d-1}$ is

$$\frac{2^{d-1}}{B(\alpha)} \prod_{i=1}^{d} x_i^{2\alpha_i - 1}.$$

In other words, the uniform distribution on the sphere arises when $\alpha_i = 1/2$, whereas $\alpha = 1$ gives the uniform distribution on the simplex (Fig. 3).

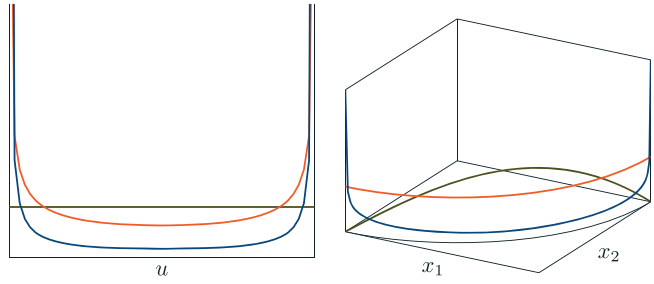

Fig. 3. Densities of different beta(α, α) distributions for $u \in (0, 1)$ (left) and their corresponding transformations to the positive quadrant of the unit circle $\mathbb{S}^1$, by the mapping $u \mapsto (\sqrt{1-u}, \sqrt{u})$ (right). $\alpha = 0.1$ (——), $\alpha = 0.5$ (——) and $\alpha = 1.0$ (——).

The area formula allows the Hausdorff measure to be easily extended to Riemannian manifolds (Federer, 1969, section 3.2.46), where

$$\mathcal{H}^m(\mathrm{d}q) = \sqrt{|G(q)|}\lambda^m(\mathrm{d}q).$$

This construction would be familiar to Bayesian statisticians as the *Jeffreys prior*, in the case where G is the Fisher–Rao metric.

When working with probability distributions on manifolds, the Hausdorff measure forms the natural reference measure and allows for reparametrization without needing to compute any additional Jacobian term. We use $\pi_{\mathcal{H}}$ to denote the density with respect to the Hausdorff measure of the distribution of interest.

Example 2.7. The von Mises distribution is a common family of distributions defined on the unit circle (Mardia and Jupp, 2000, section 3.5.4). When parametrized by an angle θ, the density with respect to the Lebesgue measure on $[0, 2\pi)$ is

$$\pi(\theta) = \frac{1}{2\pi I_0(\kappa)} \exp\{\kappa \cos(\theta - \mu)\}.$$

The embedding transformation $x = (\sin\theta, \cos\theta)$ has unit Jacobian, so the density with respect to the one-dimensional Hausdorff measure is

$$\pi_{\mathcal{H}}(x) = \frac{1}{2\pi I_0(\|c\|)} \exp\left\{c^\top x\right\},$$

where $c = (\kappa \sin\mu, \kappa\cos\mu)$ and I_k is the modified Bessel function of the first kind. In other words, it is a natural exponential family on the circle.

The von Mises–Fisher distribution is the natural extension to higher-order spheres (Mardia and Jupp, 2000, section 9.3.2) with density

$$\pi_{\mathcal{H}}(x) = \frac{\|c\|^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\|c\|)} \exp\left\{c^\top x\right\}.$$

Attempting to write this as a density with respect to the Lebesgue measure on some parametrization of the surface, such as angular coordinates, would be much more involved, as the Jacobian is no longer constant.

3. Hamiltonian Monte Carlo on embedded manifolds

RMHMC is an MCMC scheme whereby new samples are proposed by approximately solving a system of differential equations describing the paths of Hamiltonian dynamics on the manifold (Girolami and Calderhead, 2011).

The key requirement for Hamiltonian Monte Carlo is the *symplectic integrator*. This is a discretization that approximates the Hamiltonian flows yet maintains certain desirable properties of the exact solution, namely time-reversibility and volume preservation that are necessary to maintain the detailed balance conditions. The standard approach is to use a *leapfrog scheme*, which alternately updates the position and momentum via first order Euler updates (Neal, 2011).

Given a target density $\pi(q)$ (with respect to the Lebesgue measure) in some coordinate system q, RMHMC, Girolami & Calderhead (2011) utilize a Hamiltonian of the form

$$H(q, p) = -\log \pi(q) + \frac{1}{2}\log|G(q)| + \frac{1}{2}p^\top G(q)^{-1}p,$$

where G is the metric tensor. This is the negative log of the joint density (with respect to the Lebesgue measure) for (q, p), where the conditional distribution for the auxiliary momentum variable p is $\mathsf{N}(0, G(q))$.

The first two terms can be combined into the negative log of the target density with respect to the Hausdorff measure of the manifold

$$H(q, p) = -\log \pi_{\mathcal{H}}(q) + \frac{1}{2}p^\top G(q)^{-1}p. \tag{3}$$

By Hamilton's equations, the dynamics are determined by the system of differential equations

$$\frac{\mathrm{d}q}{\mathrm{d}t} = \frac{\partial H}{\partial p} = G(q)^{-1}p, \tag{4}$$

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{\partial H}{\partial q} = \nabla_q\left[\log \pi_{\mathcal{H}}(q) - \frac{1}{2}p^\top G(q)^{-1}p\right]. \tag{5}$$

As this Hamiltonian is not separable (i.e. it cannot be written as the sum of a function of q and a function of p), we are unable to apply the standard leapfrog integrator.

3.1. Geodesic integrator

Girolami & Calderhead (2011) develop a generalized leapfrog scheme, which involves composing adjoint Euler approximations to (4) and (5) in a reversible manner. Unfortunately, some of these steps do not have an explicit form and so need to be solved implicitly by fixed-point iterations. Furthermore, these updates require computation of both the inverse and derivatives of the metric tensor, which are $O(m^3)$ operations; this limits the feasibility of numerically naive implementations of this scheme for higher-dimensional problems. Finally, such a scheme assumes a global coordinate system, which may cause problems for manifolds for which none exist, such as the sphere, where artificial boundaries may be induced.

In this contribution, we instead construct an integrator by *splitting the Hamiltonian* (Hairer *et al.*, 2006, section II.5): that is, we treat each term in (3) as a distinct Hamiltonian and alternate simulating between the exact solutions.

Splitting methods have been used in other contexts to develop alternative integrators for Hamiltonian Monte Carlo (Neal, 2011, section 5.5.1) such as extending HMC to infinite-dimensional Hilbert spaces (Beskos *et al.*, 2011) and defining schemes that may reduce computational cost (Shahbaba *et al.*, 2011).

We take the first component of the splitting to be the 'potential' term

$$H^{[1]}(q, p) = -\log \pi_{\mathcal{H}}(q).$$

Hamilton's equations give the dynamics

$$\dot{q} = \frac{\partial H^{[1]}}{\partial p} = 0 \quad \text{and} \quad \dot{p} = -\frac{\partial H^{[1]}}{\partial q} = \nabla_q \log_{\mathcal{H}} \pi(q).$$

Starting at $(q(0), p(0))$, this has the exact solution

$$q(t) = q(0) \qquad \text{and} \qquad p(t) = p(0) + t\nabla_q \log \pi_{\mathcal{H}}(q)|_{q=q(0)}. \tag{6}$$

In other words, this is just a linear update to the momentum p.

The second component is the 'kinetic' term

$$H^{[2]}(q, p) = \frac{1}{2} p^\top G(q)^{-1} p. \tag{7}$$

This is simply a Hamiltonian absent of any potential term, and the solution of Hamilton's equations can be easily shown to be a geodesic flow under the Levi-Civita connection of G (Abraham and Marsden, 1978, theorem 3.7.1) or to be more precise, a co-geodesic flow $(q(t), p(t))$, where $p(t) = G(q(t))\dot{q}(t)$.

Thus, if we are able to exactly compute the geodesic flow, then we can construct an integrator by alternately simulating from the dynamics of $H^{[1]}$ and $H^{[2]}$ for some time step ϵ. Each iteration of the integrator consists of the following steps, starting at position (q, p) in the phase space:

(i) Update according to the solution to $H^{[1]}$ in (6), for a period of $\epsilon/2$ by setting

$$p \leftarrow p + \frac{\epsilon}{2} \nabla_q \log \pi_{\mathcal{H}}(q), \tag{8}$$

(ii) Update according to $H^{[2]}$, by following the geodesic flow starting at (q, p), for a period of ϵ.

(iii) Update again according to $H^{[1]}$ for a period of $\epsilon/2$ by (8).

As $H^{[1]}$ and $H^{[2]}$ are themselves Hamiltonian systems, their solutions are necessarily both reversible and symplectic. As the integrator is constructed by their symmetric composition, it will also be reversible and symplectic.

Therefore, the overall transition kernel for our Hamiltonian Monte Carlo scheme from an initial position q_0 is as follows:

(i) Propose an initial momentum p_0 from $N(0, G(q_0))$.

(ii) Map $(q_0, p_0) \mapsto (q_T, p_T)$ by running T iterations of the aforementioned integrator.

(iii) Accept the q_T as the new value with probability

$$1 \wedge \exp\{-H(q_T, p_T) + H(q_0, p_0)\}.$$

Otherwise, return the original value q_0.

As with the RMHMC algorithm, the metric G need only be known up to proportionality: scaling is equivalent to changing the time step ϵ.

3.2. Embedding coordinates

The algorithm can also be written in terms of an embedding, which avoids altogether the computation of the metric tensor and the possible lack of a global coordinate system.

Given an isometric embedding $\xi : \mathcal{M} \to \mathbb{R}^n$, then the path $x(t) = \xi(q(t))$, such that

$$\dot{x}_i(t) = \sum_j \frac{\partial x_i}{\partial q_j} \dot{q}_j(t).$$

Therefore, we can transform the phase space (q, p), where $\dot{q} = G^{-1}p$, to the embedded phase space (x, v), such that

$$v = \dot{x} = MG(q)^{-1}p = M\left(M^\top M\right)^{-1} p \quad \text{where } M_{ij} = \frac{\partial x_i}{\partial q_j},$$

because $G = M^\top M$, from (1).

By substitution, the Hamiltonian (3) can be written in terms of these coordinates as

$$H = -\log \pi_{\mathcal{H}}(x) + \tfrac{1}{2} v^\top v. \tag{9}$$

Note that the target density $\pi_{\mathcal{H}}$ is still defined with respect to the Hausdorff measure of the manifold, and so no additional log-Jacobian term is introduced.

We can rewrite the solution to $H^{[1]}$ in (6) in these coordinates. The position $x(t)$ remains constant, and by the change of variables of the operator $\nabla_q = M^\top \nabla_x$, the velocity has a linear path

$$v(t) = v(0) + tM\left(M^\top M\right)^{-1} M^\top \nabla_x \log \pi_{\mathcal{H}}(x)|_{x=x(0)}.$$

The linear operator $M\left(M^\top M\right)^{-1} M^\top$ is the 'hat matrix' from linear regression: this is the orthogonal projection onto the span of the columns of M, that is, the tangent space of the embedded manifold.

Although it is possible to compute this projection using standard least squares algorithms, it can be computationally expensive and prone to numerical instability at the boundaries of the coordinate system (e.g. at the poles of a sphere). However, for all the manifolds that we consider there exists an explicit form for an orthonormal basis N of the *normal* to the tangent space, in which case we can simply subtract the projection onto the normal:

$$v(t) = v(0) + t\left(I - NN^\top\right) \nabla_x \log \pi_{\mathcal{H}}(x)|_{x=x(0)}.$$

Finally, we require a method for sampling the initial velocity v_0. Because $p_0 \sim \mathsf{N}(0, G(q))$, it follows that

$$v_0 \sim \mathsf{N}\left(0, M\left(M^\top M\right)^{-1} M^\top\right) = \mathsf{N}\left(0, I - NN^\top\right).$$

We do not need to compute a Cholesky decomposition here: because $(I - NN^\top)$ is a projection, it is idempotent, so we can draw z from $\mathsf{N}(0, I_n)$ and project $v_0 = (I - NN^\top)z$ to obtain the necessary sample.

The resulting procedure is presented in Algorithm 1. In order to implement it for an embedded manifold $\mathcal{M} \subseteq \mathbb{R}^n$, we need to be able to evaluate the following at each $x \in \mathcal{M}$:

(i) the log-density with respect to the Hausdorff measure $\log \pi_{\mathcal{H}}$, and its gradients;
(ii) an orthogonal projection from $\mathbb{R}^n$ to the tangent space of $x \in \mathcal{M}$;
(iii) the geodesic flow from any $v \in T_x\mathcal{M}$.

Algorithm 1 The transition kernel for Hamiltonian Monte Carlo on an embedded manifold using geodesic flows.

$v \sim \mathsf{N}(0, I_n)$
$v \leftarrow v - N(x)N(x)^\top v$
$h \leftarrow \log \pi_{\mathcal{H}}(x) - \frac{1}{2} v^\top v$
$x^* \leftarrow x$
for $\tau = 1, \ldots, T$ **do**
  $v \leftarrow v + \frac{\epsilon}{2} \nabla_{x^*} \log \pi_{\mathcal{H}}(x^*)$
  $v \leftarrow v - N(x)N(x)^\top v$
  Update (x^*, v) by following the geodesic flow for a time interval of ϵ
  $v \leftarrow v + \frac{\epsilon}{2} \nabla_{x^*} \log \pi_{\mathcal{H}}(x^*)$
  $v \leftarrow v - N(x)N(x)^\top v$
end for
$h^* \leftarrow \log \pi_{\mathcal{H}}(x^*) - \frac{1}{2} v^\top v$
$u \sim \mathsf{U}(0, 1)$
if $u < \exp(h^* - h)$ **then**
  $x \leftarrow x^*$
end if

Note that by working entirely in the embedded space, we completely avoid the coordinate system q and the related problems where no single global coordinate system exists. The Riemannian metric G only appears in the Jacobian determinant term of the density: in certain examples, this can also be removed, for example by specifying the prior distribution as uniform with respect to the Hausdorff measure, as is performed in section 5.3

4. Embedded manifolds with explicit geodesics

In this section, we provide examples of embedded manifolds for which the explicit forms for the geodesic flow are known and derive the bases for the normal to the tangent space.

4.1. Affine subspaces

If the embedded manifold is flat, for example an affine subspace of $\mathbb{R}^n$, then the geodesic flows are the straight lines

$$[x(t), v(t)] = [x(0), v(0)] \begin{bmatrix} 1 & 0 \\ t & 1 \end{bmatrix}.$$

In the case of the Euclidean manifold $\mathbb{R}^n$, then the normal space to the tangent is null, and no projections are required. Hence, the algorithm reduces to the standard leapfrog scheme of HMC.

In standard HMC, it is common to utilize a 'mass' or 'preconditioning' positive-definite matrix M, in order to reduce the correlation between samples, especially where variables are highly correlated or have different scales of variation. This is directly equivalent to using the RMHMC algorithm with constant a Riemannian metric, or our geodesic procedure on the embedding of $x = L^\top q$, where L is a matrix square root such that $LL^\top = M$ (such as the Cholesky factor).

4.2. Spheres

Recall from earlier examples that the unit $(d-1)$-sphere $\mathbb{S}^{d-1}$ is an $(d-1)$-dimensional manifold embedded in $\mathbb{R}^d$, characterized by the constraint

$$x^\top x = 1,$$

with tangent space

$$\{v \in \mathbb{R}^d : x^\top v = 0\}.$$

Distributions on spheres, particularly $\mathbb{S}^1$ and $\mathbb{S}^2$, arise in many problems in directional statistics (Mardia and Jupp, 2000): examples include the von Mises–Fisher distribution (example 2.7) and the Bingham–von Mises–Fisher (BVMF) distribution (section 5.1). For many of these distributions, the normalization constants of the density functions are often computationally intensive to evaluate, which makes Monte Carlo methods particularly attractive.

As mentioned in example 2.5, the geodesics of the sphere are the great circle rotations about the origin. The geodesic flows are then

$$[x(t), v(t)] = [x(0), v(0)] \begin{bmatrix} 1 & 0 \\ 0 & \alpha^{-1} \end{bmatrix} \begin{bmatrix} \cos(\alpha t) & -\sin(\alpha t) \\ \sin(\alpha t) & \cos(\alpha t) \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \alpha \end{bmatrix}, \tag{10}$$

where $\alpha = \|v(t)\|$ is the (constant) angular velocity. The normal to the tangent space at x is x itself, so $(I - xx^\top)u$ is an orthogonal projection of an arbitrary $u \in \mathbb{R}^d$ onto the tangent space.

Other than the evaluation of the log-density and its gradient, the computations only involve vector–vector operations of addition and multiplication, so the algorithm scales linearly in d.

4.3. Stiefel manifolds

A *Stiefel manifold* $\mathbb{V}_{d,p}$ is the set of $d \times p$ matrices X such that

$$X^\top X = I.$$

In other words, it is the set of matrices with orthonormal column vectors, or equivalently, the set of p-tuples of orthogonal points in $\mathbb{S}^{d-1}$, and is a $[dp - \frac{1}{2}p(p+1)]$-dimensional manifold, embedded in $\mathbb{R}^{d \times p}$. In the special case where $d = p$, the Stiefel manifold is the *orthogonal group* $\mathbb{O}_d$: the set of $d \times d$ orthogonal matrices.

These arise in the statistical problems related to dimension reduction such as factor analysis and principal component analysis, where the aim is to find a low dimensional subspace that represents the data. They can also arise in contexts where the aim is to identify orientations, such as projections in shape analysis, or the eigendecomposition of covariance matrices.

Previously suggested methods of sampling from distributions on Stiefel manifolds, such as Hoff (2009) and Dobigeon & Tourneret (2010), have relied on column-wise Gibbs updates. Such an approach is limited to cases where the conditional distribution of the column has a conjugate form and requires the computation of an orthonormal basis for the null space of X, requiring $O(d^3)$ operations.

Again, we can find the constraints on the phase space by the time derivative of the constraint for an arbitrary curve $X(t)$ in $\mathbb{V}_{d,p}$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[X(t)^\top X(t)\right] = \dot{X}(t)^\top X(t) + X(t)^\top \dot{X}(t) = 0.$$

That is, the tangent space at X is the set

$$\left\{V \in \mathbb{R}^{d\times p} : V^\top X + X^\top V = 0\right\}.$$

If we let $\tilde{x}$ denote the matrix X written as a vector in $\mathbb{R}^{dp}$ by stacking the columns $x_1, \ldots, x_p$, then an orthonormal basis N for the normal to the tangent space has p vectors of the form

$$\begin{bmatrix} x_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ x_2 \\ \vdots \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ 0 \\ \vdots \\ x_p \end{bmatrix}$$

and $\binom{p}{2}$ vectors of the form

$$\begin{bmatrix} \frac{1}{\sqrt{2}}x_2 \\ \frac{1}{\sqrt{2}}x_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{1}{\sqrt{2}}x_3 \\ 0 \\ \frac{1}{\sqrt{2}}x_1 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \frac{1}{\sqrt{2}}x_3 \\ \frac{1}{\sqrt{2}}x_2 \\ \vdots \\ 0 \end{bmatrix} \ldots$$

For an arbitrary vector $\tilde{u} \in \mathbb{R}^{dp}$, the projection onto the tangent space is then

$$\tilde{u} - NN^\top \tilde{u} = \begin{bmatrix} u_1 - x_1\left(x_1^\top u_1\right) - \frac{1}{2}x_2\left(x_1^\top u_2 + x_2^\top u_1\right) - \ldots \\ u_2 - x_2\left(x_2^\top u_2\right) - \frac{1}{2}x_1\left(x_2^\top u_1 + x_1^\top u_2\right) - \ldots \\ \vdots \end{bmatrix}.$$

This can be more easily written in matrix form: for an arbitrary $U \in \mathbb{R}^{d\times p}$, the orthogonal projection onto the Stiefel manifold is

$$U - \frac{1}{2}X\left(X^\top U + U^\top X\right).$$

The geodesic flows are more complicated than the spherical case. For $p > 1$, they are no longer simple rotations but can be expressed in terms of matrix exponentials (Edelman *et al.*, 1999, page 310)

$$[X(t), V(t)] = [X(0), V(0)] \exp\left\{t\begin{bmatrix} A & -S(0) \\ I & A \end{bmatrix}\right\}\begin{bmatrix} \exp\{-tA\} & 0 \\ 0 & \exp\{-tA\} \end{bmatrix},$$

where $A = X(t)^\top V(t)$ is a skew-symmetric matrix that is constant over the geodesic, and $S(t) = V(t)^\top V(t)$ is non-negative definite.

Although matrix exponentials can be quite computationally expensive, we note that the largest exponential of these is of a $2p \times 2p$ matrix, which requires $O(p^3)$ operations. Other than this and the evaluations of the log-density and its gradients, all the other operations are simple matrix additions and multiplications, the largest of which can be performed in $O(dp^2)$ operations; hence, the algorithm scales linearly with d.

For the orthogonal group $\mathbb{O}_d$, the geodesics have the simpler form (Edelman *et al.*, 1999, equation 2.14)

$$[X(t), V(t)] = [X(0), V(0)]\begin{bmatrix} \exp\{tA\} & 0 \\ 0 & \exp\{tA\} \end{bmatrix}.$$

As A is skew-symmetric, Rodrigues' formula gives an explicit form of $\exp\{tA\}$ when $d = 3$ in terms of simple trigonometric functions, and this can be extended into higher dimensions (Gallier and Xu, 2002; Cardoso and Leite, 2010).

4.4. Product manifolds

Given two manifolds $\mathcal{M}_1$ and $\mathcal{M}_2$, their Cartesian product

$$\mathcal{M}_1 \times \mathcal{M}_2 = \{(x_1, x_2) : x_1 \in \mathcal{M}_1, x_2 \in \mathcal{M}_2\}$$

is also a manifold.

Product manifolds arise naturally in many statistical problems; for example, extensions of the von Mises distributions to $\mathbb{S}^1 \times \mathbb{S}^1$ (a torus) have been used to model molecular angles (Singh *et al.*, 2002), and the network eigenmodel in section 5.3 has a posterior distribution on $\mathbb{V}_{m,p} \times \mathbb{R}^p \times \mathbb{R}$.

The geodesics of a product manifold are of the form (γ_1, γ_2), where each γ_i is a geodesic of $\mathcal{M}_i$. Likewise, the tangent vectors are of the form (v_1, v_2), where each v_i is a tangent to $\mathcal{M}_i$. Consequently, for an arbitrary vector (u_1, u_2), the orthogonal projection onto the tangent space is $\left(\left[I - N_1 N_1^\top\right] u_1, \left[I - N_2 N_2^\top\right] u_2\right)$, where N_i is an orthonormal basis of $\mathcal{M}_i$.

As a result, when implementing our geodesic Monte Carlo scheme on a product manifold, the key operations (addition of gradient, projection and geodesic update) can be essentially performed in parallel, the only operations requiring knowledge of the other variables being the computation of the log-density and its gradient. Moreover, when tuning the algorithm, it is possible to choose different ϵ values for each constituent manifold, which can be helpful when variables have different scales of variation.

5. Illustrative examples

5.1. Bingham–von Mises–Fisher distribution

The BVMF distribution is the exponential family on $\mathbb{S}^{d-1}$ with linear and quadratic terms, with density of the form

$$\pi_{\mathcal{H}}(x) \propto \exp\left\{c^\top x + x^\top A x\right\},$$

where c is a vector of length d, and A is a $d \times d$ symmetric matrix (Mardia and Jupp, 2000, section 9.3.3).

The Bingham distribution arises as the special case where $c = 0$: this is an axially bimodal distribution, with the modes corresponding to the eigenvector of the largest eigenvalue. The BVMF distribution may or may not be bimodal, depending on the parameter values.

Hoff (2009) develops a Gibbs-style method for sampling from BVMF distribution by first transforming $y = E^\top x$, where $E^\top \Lambda E$ is the eigendecomposition of A. Each element y_i of y is updated in random order, conditional $u \in \mathbb{S}^{d-2}$, where $u_j = y_j / \sqrt{1 - y_i^2}$ for $j \neq i$. The $y_i \mid u$ are sampled using a rejection sampling scheme with a beta envelope; however, as noted by Brubaker *et al.* (2012), this can give exponentially poor acceptance probabilities (of the order of 10^{-100}) for certain parameter values, particularly when c is large in the direction of the negative eigenspectra.

Implementing our geodesic sampling scheme for the BVMF distribution is straightforward, as the gradient of the log-density is simply $c + 2Ax$, and extremely fast to run, with run times that are independent of the parameter values. However, as with any gradient-based method, it has difficulty switching between multiple modes (Fig. 4).

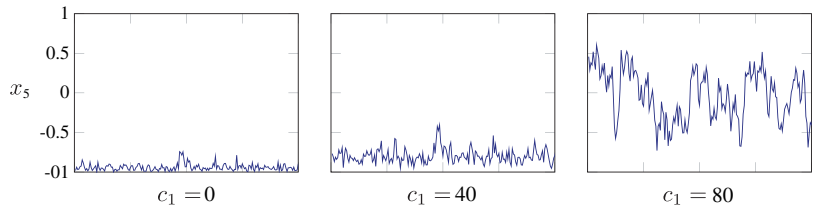

Fig. 4. Trace plots of x_5 from 200 samples from the spherical geodesic Monte Carlo sampler (with parameters $\epsilon = 0.01, T = 20$) for a Bingham–von Mises–Fisher distribution, with parameters $A = \text{diag}(-20, -10, 0, 10, 20)$ and $c = (c_1, 0, 0, 0, 0)$. When the distribution is bimodal ($c_1 = 0, 40$), the sampler has difficulty moving between the modes.

A common method of alleviating this problem is to utilize tempering schemes (Neal, 2011, section 5.5.7): these operate by sampling from a class of 'higher temperature' distributions with densities of the form

$$[\pi_{\mathcal{H}}(x)]^{\rho} \quad \text{where } 0 \leq \rho \leq 1.$$

Note that this constitutes a simple linear scaling of the log-density and so can be easily incorporated into our method. *Parallel tempering* (Geyer, 1991; Liu, 2008, section 10.4) utilizes multiple chains, each targeting a density with a different temperature. The scheme operates by alternately updating the individual chains, which can be performed in parallel, and randomly switching the values of neighbouring chains with a Metropolis–Hastings correction to maintain detailed balance. The results of utilizing such a scheme are shown in Fig. 5

5.2. *Non-conjugate simplex models*

We can use the transformation to the sphere to sample from distributions on the simplex Δ^{d-1}. These arise in many contexts, particularly as prior and posterior distributions for discrete-valued random variables such as the multinomial distribution.

If each observation x from the multinomial is completely observed, then the contribution to the likelihood is then θ_x, giving a full likelihood of at most d terms of form

$$L(\theta) = \prod_{i=1}^{d} \theta_i^{N_i},$$

which is conjugate to a Dirichlet prior distribution.

Complications arise if observations are only partially observed. For example, we may have *marginal* observations, which are only observed to a set S, in which case the likelihood term is $\sum_{s \in S} \theta_s$, or *conditional* observations, where the sampling was constrained to occur within

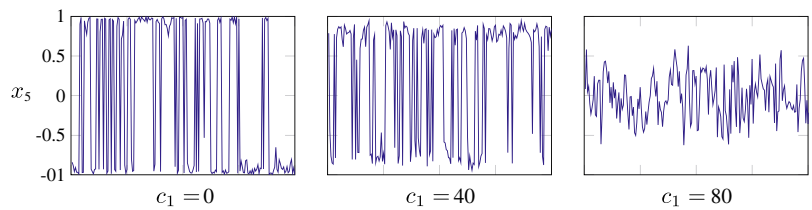

Fig. 5. Trace plots of a simulated tempering scheme applied to the target of Fig. 4, using 10 parallel chains to transition between multiple modes. The values of ρ were $0.1, 0.2, \ldots, 1.0$, and 10 random exchanges were applied between parallel geodesic Monte Carlo updates.

a set T, with likelihood term $\theta_x/(\sum_{t\in T}\theta_t)$. These terms destroy the conjugacy and make computation very difficult.

Such models arise under a *case-cohort design* (Le Polain deWaroux *et al.*, 2012), the risk factors of a particular disease: for the case sample, the risk factors are observed conditional on the person having the disease, and for the cohort sample, the risk factors are observed marginally (as disease status is unknown). Overall population statistics may provide some further information as to the marginal probability of the disease.

The `hyperdirichlet` R package (Hankin, 2010) provides an interface and examples for dealing with this type of data. We consider the volleyball data from this package: the data arise from a sports league for nine players, where each match consists of two disjoint teams of players, one of which is the winner. The probability of a team T_1 beating T_2 is assumed to be

$$\frac{\sum_{t\in T_1} p_t}{\sum_{t\in T_1\cup T_2} p_t},$$

where $p = (p_1,\ldots,p_9) \in \Delta^8$. We compare three different methods in sampling from the posterior distribution for p under a Dirichlet $(\alpha\mathbf{1})$ prior, for different values of α. The results are presented in Table 1.

The first is a simple random-walk Metropolis–Hastings algorithm. To ensure that the planar constraint $\sum_{i=1}^9 p_i = 1$ is satisfied, the proposals are made from a degenerate $\mathsf{N}(x, \epsilon^2[I - nn^\top])$, where $n = d^{-1/2}\mathbf{1}$ is the normal to the simplex.

The second is a random walk on the sphere, based on the square root transformation to the sphere from example 2.4, using proposals of the form

$$x_{\text{proposed}} = x\cos(\|\delta\|) + \frac{\delta}{\|\delta\|}\sin(\|\delta\|), \quad \text{where } \delta \sim \mathsf{N}\left(0, \epsilon^2\left[I - xx^\top\right]\right).$$

Although this only defines the distribution on the positive orthant, we can extend this distribution to the entire sphere by reflecting about the axes (because we only require knowledge of the density up to proportionality, we can ignore the fact that it is now 2^d times larger). One benefit of this transformation is that the surface is now smooth and without boundaries, so proposals outside the positive orthant can be accepted.

The third is the geodesic Monte Carlo algorithm on the simplex. We can ensure that the planar constraint is satisfied via the affine constraint methods in 4.1; however, we need to further ensure that the integration paths satisfy the positivity constraints, which can be achieved by reflecting the path whenever it violates the constraint (see the Appendix for further details).

The fourth is our proposed geodesic scheme based on the spherical transformation. As the integrator does not pass any boundaries, no reflections are required.

Table 1. *Average effective sample size (ESS) across coordinates per 100 samples, and per second, of the Volleyball model under a Dirichlet* $(\boldsymbol{\alpha 1})$ *prior from 1 000 000 samples. For all samplers,* $\epsilon = 0.01$*, and for the HMC algorithms,* $T = 20$ *integration steps were used. We attempted some tuning of the parameters but were unable to obtain any noticeable changes in performance*

	$\alpha = 0.1$		$\alpha = 0.5$		$\alpha = 1.0$		$\alpha = 5.0$	
	ESS %	ESS/ second	ESS %	ESS/ second	ESS %	ESS/ second	ESS %	ESS/ second
RW-MH	0.0064	6.10	0.113	71.1	0.36	158	0.84	290
Spherical RW	0.0089	2.48	0.143	37.6	0.19	51	0.45	123
Simplex HMC	0.0034	0.0079	0.037	0.12	53.4	611	75.6	976
Spherical HMC	0.0187	0.327	77.3	1374	92.6	1616	187.4	3262

For small values of α, both geodesic Monte Carlo samplers perform poorly, due to the concentration of the density at the boundaries. These peaks cause particular problems for the Hamiltonian-type algorithms, as the discontinuous gradients mean that the integration paths give poor approximations to the true Hamiltonian paths, resulting in poor acceptance probabilities. Moreover, for the algorithm on the simplex, the frequent reflections add to the computational cost.

However, when $\alpha = 0.5$, the spherical geodesic sampler improves markedly: recall from example 2.6 and Fig. 3 that the Dirichlet (0.5) prior is uniform on the sphere, giving continuous gradients. On the simplex, however, the density remains peaked at the boundaries. The simplex sampler improves considerably for values of $\alpha \geq 1$ (where the gradient is now flat or negative); however the spherical algorithm still retains a slight edge. Interestingly, the spherical random walk sampler performs poorly in all of the examples.

5.3. *Eigenmodel for network data*

We use the network eigenmodel of Hoff (2009) to demonstrate how Stiefel manifold models can be used for dimension reduction and how our geodesic sampling scheme may be used for Stiefel and product manifolds. This is a model for a graph on a set of m nodes, where for each unordered pair of nodes $\{i, j\}$, there is a binary observation $Y_{\{i,j\}}$ indicating the existence of an edge between i and j.

The specific example of Hoff (2009) is a protein interaction network, where for $m = 270$ proteins, the existence of the edge indicates whether or not the pair of proteins interact.

The model represents the network by assuming a low ($p = 3$) dimensional representation for the probability of an edge

$$P(Y_{\{i,j\}} = 1) = \Phi\left([U\Lambda U^\top]_{ij} + c\right),$$

where $\Phi : \mathbb{R} \to (0, 1)$ is the probit link function, U is an orthonormal $m \times p$ matrix and Λ is a $p \times p$ diagonal matrix. U is assumed to have a uniform prior distribution on $\mathbb{V}_{m,p}$ (with respect to the Hausdorff measure), the diagonal elements of Λ have a $\mathsf{N}(0, m)$ distribution and $c \sim \mathsf{N}(0, 10^2)$.

Hoff (2009) uses column-wise Gibbs updates for sampling U, exploiting the fact that the probit link provides an augmentation that allows these to be sampled as a BVMF distribution. However, as mentioned in section 4.3, this requires computing the full null space of U at each iteration.

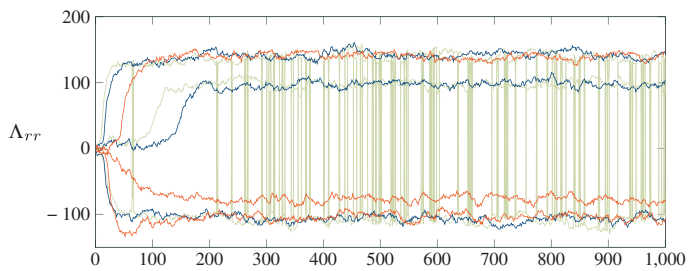

Fig. 6. Trace plots of 1000 samples of the diagonal elements of Λ from geodesic Monte Carlo sampler on the network eigenmodel. One chain (——) converges to the same mode as Hoff (2009), while the other (——) converges to a local mode, with approximately 10^{-36} of the density. By incorporating this into a parallel tempering scheme (——), the sampler rapidly finds the higher mode and is able to switch between the various permutations.

We implement geodesic Monte Carlo on the product manifold, details of which are given in the Appendix. Trace plots from two chains of the diagonal elements of Λ appear in Fig. 6: note that one chain appears to get stuck in a local mode, while the other converges to the same as the method of Hoff (2009).

By incorporating this approach into a parallel tempering scheme, the model is able to find the larger mode with greater reliability. Moreover, unlike the algorithm of Hoff (2009), it is capable of switching between the permutations of Λ, which would further suggest that this is indeed the global mode.

6. Conclusion and discussion

We have presented a scheme for sampling from distributions defined on manifolds embedded in Euclidean space by exploiting their known geodesic structure. This method has been illustrated using applications from directional statistics, discrete data analysis and network analysis. This method does not require any conjugacy, allowing greater flexibility in the choice of models: for instance, it would be straightforward to change the probit link in section 5.3 to a logit. Moreover, when used in conjunction with a tempering scheme, it is capable of efficiently exploring complicated multimodal distributions.

Our approach could be widely applicable to problems in directional statistics, such as the estimation of normalization constants that are often otherwise numerically intractable. The method of transforming the simplex to the sphere could be useful for applications dealing with high-dimensional discrete data, such as statistical genetics and language modelling. Stiefel manifolds arise naturally in dimension reduction problems, and our methods could be particularly useful where the data are not normally distributed, for instance the analysis of survey data with discrete responses, such as Likert scale data. Furthermore, this method could be utilized in statistical shape and image analysis for determining the orientation of objects in projected images.

The major constraint of this technique is the requirement of an explicit form for the geodesic flows that can be easily evaluated numerically. These are not often available; for instance, the geodesic paths of ellipsoids require often computationally intensive elliptic integrals.

Of the examples we consider, the geodesics of the Stiefel manifold case are the most demanding, due to the matrix exponential terms. An alternative approach would be to utilize a Metropolis-within-Gibbs style scheme over subsets of columns, for example by updating a pair of columns such that they remain orthogonal to the remaining columns.

However, once the geodesics and the orthogonal tangent projection of the manifold are known, the remaining process of computing the derivatives is straightforward, and could be easily implemented using automatic differentiation tools, as is used in the `Stan` MCMC library (Stan Development Team, 2012), currently under development.

Acknowledgements

Simon Byrne is funded by a BBSRC grant (BB/G006997) and an EPSRC Postdoctoral Fellowship (EP/K005723). Mark Girolami is funded by an EPSRC Established Career Fellowship (EP/J016934) and a Royal Society Wolfson Research Merit Award.

Appendix A: Reflecting at boundaries of the simplex

Neal (2011, section 5.5.1.5) notes that when an integration path crosses a boundary of the sample space, it can be reflected about the normal to the boundary to keep it within the desired

space. He considers boundaries that are orthogonal to the ith axis, with normals of the form δ_i. Whenever such a constraint is violated, the position and velocity are replaced by

$$x_i' = b_i + (b_i - x_i) \quad v_i' = -v_i,$$

where b_i is the boundary (either upper or lower). As no other coordinates are involved in this reflection, this can be performed in parallel for all constrained coordinates.

However, for the simplex Δ^{d-1}, the normals are not of the form δ_i, as this would result in the path being reflected off the plane $\{x : \sum_i x_i = 1\}$. Instead, we need to reflect about the projection of δ_i onto the plane, that is,

$$\tilde{n}_i = \frac{\tilde{m}_i}{\|\tilde{m}_i\|} = \frac{d\delta_i - \mathbf{1}}{\sqrt{d(d-1)}} \quad \text{where} \quad \tilde{m}_i = \delta_i - \left(d^{-1/2}\mathbf{1}\right)\left(d^{-1/2}\mathbf{1}\right)^\top \delta_i.$$

A procedure for performing the position updates is given in Algorithm 2.

Algorithm 2 The geodesic updates on the simplex incorporating reflection off the boundaries.

$\omega \leftarrow \epsilon$
while $\omega > 0$ **do**
  $(\kappa, j) \leftarrow (\min, \arg\min_i)\{-x_i/v_i : v_i < 0\}$ {The time until any coordinate is negative-valued: this can only occur when the velocity is negative.}
  $x \leftarrow x + \min(\omega, \kappa)v$
  $\omega \leftarrow \omega - \min(\omega, \kappa)$
  if $\omega > 0$ **then**
    $v \leftarrow v - 2\tilde{n}_j\tilde{n}_j^\top v$
  end if
end while

Unfortunately, this procedure cannot be applied to the RMHMC integrator proposed by Girolami & Calderhead (2011), as the implicit steps involved make it difficult to calculate the reflections.

Appendix B: Network eigenmodel

Define the $p \times p$ symmetric matrices $\eta = U\Lambda U^\top + c$ and Y^*, where

$$Y_{ij}^* = \begin{cases} 1 & Y_{\{i,j\}} = 1 \\ 0 & i = j \\ -1 & Y_{\{i,j\}} = 0 \end{cases}.$$

Then using the property that $1 - \Phi(x) = \Phi(-x)$, the log-density of the posterior is

$$\log \pi_{\mathcal{H}}(U, \Lambda, c) = \sum_{\{i,j\}} \log \Phi\left(Y_{ij}^* \eta_{ij}\right) - \sum_{r=1}^{p} \frac{\Lambda_{rr}^2}{2m} - \frac{c^2}{200} + \text{constant}.$$

The gradients with respect to the parameters are

$$\frac{\partial \log \pi_{\mathcal{H}}}{\partial U_{ir}} = \sum_{j=1}^{m} \frac{\partial \log \pi_{\mathcal{H}}}{\partial \eta_{ij}} U_{jr} \Lambda_{rr},$$
$$\frac{\partial \log \pi_{\mathcal{H}}}{\partial \Lambda_{rr}} = \sum_{\{i,j\}} \frac{\partial \log \pi_{\mathcal{H}}}{\partial \eta_{ij}} U_{ir} U_{jr} - \frac{\Lambda_{rr}}{m},$$
$$\frac{\partial \log \pi_{\mathcal{H}}}{\partial c} = \sum_{\{i,j\}} \frac{\partial \log \pi_{\mathcal{H}}}{\partial \eta_{ij}} - \frac{c}{100},$$

where the gradients with respect to the linear predictors are

$$\frac{\partial \log \pi_{\mathcal{H}}}{\partial \eta_{ij}} = Y^*_{ij} \frac{\phi\left(Y^*_{ij}\eta_{ij}\right)}{\Phi\left(Y^*_{ij}\eta_{ij}\right)}.$$

The programme was implemented in MATLAB (The MathWorks Inc., Natick, Massachusetts, USA). To avoid numerical overflow errors, the ratio $\phi(x)/\Phi(x)$, as well as $\log \Phi(x)$ for negative values of x, are calculated using the `erfcx` function. The matrix exponential terms were calculated using the inbuilt `expm` function, which utilizes a Padé approximation with scaling and squaring.

Different ϵ values were used for each parameter: $\epsilon_U = 0.005$, $\epsilon_\Lambda = 0.1$ and $\epsilon_c = 0.001$. $T = 20$ integration steps were run for each iteration. The parallel tempered version utilized 20 parallel chains, with 10 proposed exchanges between parallel updates.

References

Abraham, R. & Marsden, J. E. (1978). *Foundations of mechanics*, (2nd ed.)., Benjamin/Cummings Publishing Co. Inc. Advanced Book Program, Reading, Mass. ISBN: 0-8053-0102-X.

Amari, S. & Nagaoka, H. (2000). *Methods of information geometry*, Translations of Mathematical Monographs, vol. 191, American Mathematical Society, Providence, RI. ISBN: 0-8218-0531-2.

Beskos, A., Pinski, F. J., Sanz-Serna, J. M. & Stuart, A. M. (2011). Hybrid Monte Carlo on Hilbert spaces. *Stochastic Process. Appl.* **121**, (10), 2201–2230, DOI 10.1016/j.spa.2011.06.003.

Brubaker, M., Salzmann, M. & Urtasun, R. (2012). A family of MCMC methods on implicitly defined manifolds. In *JMLR Workshop and Conference Proceedings*, Vol. 22; 161–172. Available on http://jmlr.csail.mit.edu/proceedings/papers/v22/brubaker12/brubaker12. pdf.

Cardoso, J. R. & Leite, F. S. (2010). Exponentials of skew-symmetric matrices and logarithms of orthogonal matrices. *J. Comput. Appl. Math.* **233**, (11), 2867–2875. DOI: 10.1016/j.cam.2009.11.032.

Diaconis, P., Holmes, S. & Shahshahani, M. (2013). Sampling from a manifold. In *Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton* (eds G. Jones & X. Shen), Institute of Mathematical Statistics.

do Carmo, M. P. (1976). *Differential geometry of curves and surfaces*, Prentice-Hall Inc., Englewood Cliffs, N.J.

do Carmo, M. P. (1992). *Riemannian geometry*, Mathematics: theory & applications, Birkhäuser Boston Inc., Boston, MA. ISBN: 0-8176-3490-8.

Dobigeon, N. & Tourneret, J.-Y. (2010). Bayesian orthogonal component analysis for sparse representation. *IEEE Trans. Signal Process.* **58**, (5), 2675–2685. ISSN: 1053-587X, DOI: 10.1109/TSP.2010.2041594, Available on http://dx.doi.org/10.1109/TSP.2010.2041594.

Duane, S., Kennedy, A. D., Pendleton, B. J. & Roweth, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B* **195**, 216–222.

Edelman, A., Arias, T. A. & Smith, S. T. (1999). The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **20**, (2), 303–353. DOI: 10.1137/S0895479895290954.

Federer, H. (1969). *Geometric measure theory*, Die Grundlehren der mathematischen Wissenschaften, Band 153, Springer-Verlag New York Inc., New York.

Gallier, J. & Xu, D. (2002). Computing exponentials of skew-symmetric matrices and logarithms of orthogonal matrices. *Int. J. Rob. Autom.* **17**, (4), 1–11.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: The 23rd Symposium on the Interface* (ed Keramigas, E.), Interface Foundation, Fairfax; 156–163.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (eds). (1996). *Markov chain Monte Carlo in practice*, Interdisciplinary Statistics, Chapman & Hall, London. ISBN: 0-412-05551-1.

Girolami, M. & Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73**, (2), 123–214. With discussion and a reply by the authors, DOI: 10.1111/j.1467-9868.2010.00765.x.

Hairer, E., Lubich, C. & Wanner, G. (2006). *Geometric numerical integration*, (2nd ed.)., Springer Series in Computational Mathematics, vol. 31, Springer-Verlag, Berlin. Structure-preserving algorithms for ordinary differential equations, ISBN: 3-540-30663-3; 978-3-540-30663-4.

Hankin, R. K. S. (2010). A generalization of the Dirichlet distribution. *J. Stat. Softw.* **33**, (11), 1–18. Available on http://www.jstatsoft.org/v33/i11.

Hoff, P. D. (2009). Simulation of the matrix Bingham-von Mises-Fisher distribution, with applications to multivariate and relational data. *J. Comput. Graph. Statist.* **18**, (2), 438–456. DOI: 10.1198/jcgs.2009.07177.

Jolliffe, I. T. (1986). *Principal component analysis*, Springer Series in Statistics, Springer-Verlag, New York. ISBN: 0-387-96269-7.

Konukoglu, E., Relan, J., Cilingir, U., Menze, B. H., Chinchapatnam, P., Jadidi, A., Cochet, H., Hocini, M., Delingette, H., Jaïs, P., Haïssaguerre, M., Ayache, N. & Sermesant, M. (2011). Efficient probabilistic model personalization integrating uncertainty on data and parameters: application to eikonal-diffusion models in cardiac electrophysiology. *Prog. Biophys. Mol. Biol.* **107**, (1), 134–146.

Le Polain de Waroux, O., Maguire, H. & Moren, A. (2012). The case-cohort design in outbreak investigations. *Euro Surveillance: Bulletin Europeen sur les Maladies Transmissibles* **17**, (25), 11–15.

Liu, J. S. (2008). *Monte Carlo strategies in scientific computing*, Springer Series in Statistics, Springer, New York. pp. xvi+343. isbn: 978-0-387-76369-9; 0-387-95230-6.

Mardia, K. V. & Jupp, P. E. (2000). *Directional statistics*, Wiley Series in Probability and Statistics, John Wiley & Sons Ltd., Chichester. ISBN: 0-471-95333-4.

Martin, J., Wilcox, L. C., Burstedde, C. & Ghattas, O. (2012). A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM J. Sci. Comput.* **34**, (3), A1460–A1487.

Morgan, F. (2009). *Geometric measure theory*, (4th ed.)., Elsevier/Academic Press, Amsterdam. A beginner's guide, ISBN: 978-0-12-374444-9.

Nash, J. (1956). The imbedding problem for Riemannian manifolds. *Ann. of Math. (2)* **63**, 20–63.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov chain Monte Carlo*, Chapman & Hall/CRC Handb. Mod. Stat. Methods CRC Press, Boca Raton, FL; 113–162.

Raue, A., Kreutz, C., Theis, F. J. & Timmer, J. (2012). Joining forces of Bayesian and frequentist methodology: a study for inference in the presence of non-identifiability. *Phil. Trans. R. Soc. A* **371**, (1984).

Shahbaba, B., Lan, S., Johnson, W. O. & Neal, R. M. (2011). *Split Hamiltonian Monte Carlo*. arXiv: 1106.5941.

Singh, H., Hnizdo, V. & Demchuk, E. (2002). Probabilistic model for two dependent circular variables. *Biometrika* **89**, (3), 719–723. DOI: 10.1093/biomet/89.3.719.

Stan Development Team. (2012). *Stan: a C++ library for probability and sampling, version 1.0*. Available on http://mc-stan.org/.

Vanlier, J., Tiemann, C. A., Hilbers, P. A. J. & van Riel, N. A. W. (2012). An integrated strategy for prediction uncertainty analysis. *Bioinformatics* **28**, (8), 1130–1135.

Received January 2013, in final form June 2013

Simon Byrne, Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK.
E-mail: simon.byrne@ucl.ac.uk