ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Geometry and Dynamics for Markov Chain Monte Carlo

Alessandro Barp,[1,2] François-Xavier Briol,[1,2,3] Anthony D. Kennedy,[2,4] and Mark Girolami[1,2]

[1]Department of Mathematics, Imperial College London, London SW7 2AZ, United Kingdom; email: a.barp16@imperial.ac.uk, m.girolami@imperial.ac.uk

[2]The Alan Turing Institute, British Library, London NW1 2DB, United Kingdom

[3]Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom; email: f-x.briol@warwick.ac.uk

[4]School of Physics and Astronomy, University of Edinburgh, Edinburgh EH9 3JZ, United Kingdom; email: adk@ph.ed.ac.uk

Annu. Rev. Stat. Appl. 2018. 5:451–71

First published as a Review in Advance on December 8, 2017

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-031017-100141

Keywords

Markov chain Monte Carlo, information geometry, Hamiltonian mechanics, symplectic integrators, shadow Hamiltonians

Abstract

Markov chain Monte Carlo methods have revolutionized mathematical computation and enabled statistical inference within many previously intractable models. In this context, Hamiltonian dynamics have been proposed as an efficient way of building chains that can explore probability densities efficiently. The method emerges from physics and geometry, and these links have been extensively studied over the past thirty years. The aim of this review is to provide a comprehensive introduction to the geometric tools used in Hamiltonian Monte Carlo at a level accessible to statisticians, machine learners, and other users of the methodology with only a basic understanding of Monte Carlo methods. This will be complemented with some discussion of the most recent advances in the field, which we believe will become increasingly relevant to scientists.

1. INTRODUCTION

1.1. Markov Chain Monte Carlo

One aim of Monte Carlo methods is to sample from a target distribution, that is, to generate a set of independent and identically distributed (i.i.d.) samples $x^{(i)}$ with respect to the density π of this distribution. Sampling from such a distribution enables the estimation of the integral $\mathbb{E}_\pi[f] = \int_{\mathcal{X}} f \mathrm{d}\Pi$ of a function $f : \mathcal{X} \to \mathbb{R}$ with respect to its corresponding probability measure Π by $\frac{1}{m}\sum_i^m f\left(x^{(i)}\right)$. Formally, the target density is a nonnegative (almost everywhere) measurable function $\pi : \mathcal{X} \to \mathbb{R}^+$, where $\mathcal{X} \subset \mathbb{R}^d$ is the sample space of a measurable space with Lebesgue measure μ, corresponding to the probability measure $\Pi(A) = \int_A \pi \mathrm{d}\mu$.

Often we only know π up to a multiplicative constant, that is, we are able to evaluate $\tilde{\pi}$ where $\pi = \tilde{\pi}/Z$ for some $Z \in \mathbb{R}^+$. For example, this is the case in Bayesian statistics, where the normalization constant Z is the model evidence, which is itself a complicated integral not always available in closed form. Even when we know the value of Z, sampling from π is challenging, particularly in high dimensions where high-probability regions are usually concentrated on small subsets of the sample space (MacKay 2003). There are only a few densities for which we can easily generate samples.

The first Markov chain Monte Carlo (MCMC) algorithm appeared in physics (Metropolis et al. 1953) as a way of tackling these issues. The problem investigated was a large system of particles, and the aim was to compute the expected value of physical quantities. The high dimension of the system made it impossible to use numerical methods or standard Monte Carlo to compute the integral. Instead, they proposed a method based on generating samples from an arbitrary random walk and adding an accept/reject step to ensure these samples originate from the correct distribution. Despite extensive use in statistical mechanics and spatial statistics, MCMC remained unknown to the mainstream statistical literature during the following twenty years (Hastings 1970). Gelfand & Smith (1990) made the connection to more classical problems and brought MCMC to a wider public, thus marking the beginning of the MCMC revolution in statistics (Robert & Casella 2011).

The idea behind MCMC methods (Meyn & Tweedie 1993, Robert & Casella 2004) is to generate approximately i.i.d. samples from the target π by constructing a Markov chain whose stationary density is π and using samples from its path. Recall that a Markov chain is a sequence of random variables $(X_0, X_1, \ldots)$ such that the distribution of X_r depends only on X_{r-1}. A Markov chain may be specified by an initial density $h_0(x)$ for X_0 and a density $T(x' \leftarrow x)$ from which we can sample. The density of X_r is then defined by $h_r(x') = \int T(x' \leftarrow x) h_{r-1}(x) \mathrm{d}x$. The density π is called a stationary density of the Markov chain if, whenever $X_r \sim \pi$, then $X_{r+1} \sim \pi$, or in other words, $\pi(x') = \int T(x' \leftarrow x)\pi(x)\mathrm{d}x$. If the Markov chain is ergodic, it will converge to its stationary distribution independently of its initial distribution. A common way to guarantee that π is indeed the invariant density of the chain (which then asymptotically generates samples from π) is to demand that it satisfies the detailed balance condition $\pi(x)T(x' \leftarrow x) = \pi(x')T(x \leftarrow x')$. Intuitively, this condition requires that the probabilities of moving from state x to x' and from x' to x are equal. Detailed balance is sufficient but not necessary for invariance with respect to π (Diaconis et al. 2000).

The Metropolis–Hastings algorithm constructs a Markov chain converging to the desired target π by the means of a proposal kernel P, where for each $x \in \mathcal{X}$, $P(\cdot, x)$ is a density on the state space from which we can sample. Given the current state x_r,

1. Propose a new state $y \sim P(\cdot, x_r)$.
2. Accept y with probability $A(y|x_r) := \min\left\{1, \frac{\pi(y)P(x_r,y)}{\pi(x_r)P(y,x_r)}\right\}$, else set $x_{r+1} = x_r$.

This induces a transition density $T(y \leftarrow x) := P(y,x)A(y|x_r) + \mathbf{1}_{\{y=x_r\}}(1 - A(y|x_r))$, where $\mathbf{1}_{\{y=x_r\}} : \mathcal{X} \to \{0,1\}$ takes value 1 when $y = x_r$ and 0 otherwise. This quantity does not rely on the normalization constant Z, which cancels out in the ratio.

1.2. Motivation for the Use of Geometry

In principle, there are only mild requirements on the proposal P to obtain an asymptotically correct algorithm; however, the choice of P will be very significant for the performance of the algorithm. Intuitively, the aim is to choose a proposal that will favor values with a high probability of acceptance while also exploring the state space well (i.e., have small correlations with the current state). A common choice is a symmetric density (e.g., Gaussian) centered on the current state of the chain, leading to the random-walk Metropolis (RWM) algorithm. A more advanced algorithm is the Metropolis-adjusted Langevin algorithm (MALA) (Rossky et al. 1978, Scalettar et al. 1986, Roberts & Rosenthal 1998), which uses the path of a diffusion that is invariant to the target distribution.

Concentration of measure is a well-known phenomenon in high dimensions (Ledoux 2001) and is linked to concentration of volume (also commonly referred to as the curse of dimensionality). An intuitive example, often used to describe this phenomenon, is that of a sphere S^d embedded in the unit cube. Most of the volume of the cube lies outside the sphere, and this is increasingly the case for higher d. Similarly, probability measures will tend to concentrate around their mean in high dimensions (MacKay 2003, Betancourt 2017), making the use of RWM inefficient, since it does not adapt to the target distribution.

To avoid issues with high curvature and concentration of measure, Duane et al. (1987) proposed a method based on approximate simulation of a Hamiltonian dynamical system with potential energy given by the log-target density. Informally, this has the advantage of directing the Markov chain toward areas of high probability and hence providing more efficient proposals (see **Figure 1*a***). This method was originally named hybrid Monte Carlo, as it was a hybrid of molecular dynamics (microcanonical) and momentum heat bath (Gibbs sampler). The method is now also commonly known as Hamiltonian Monte Carlo (HMC) (Neal 2011).

HMC has been used throughout statistics but has also spanned a wide range of fields including biology (Berne & Straub 1997, Hansmann & Okamoto 1999, Kramer et al. 2014), medicine (Konukoglu et al. 2011, Schroeder et al. 2013), computer vision (Choo & Fleet 2001), chemistry

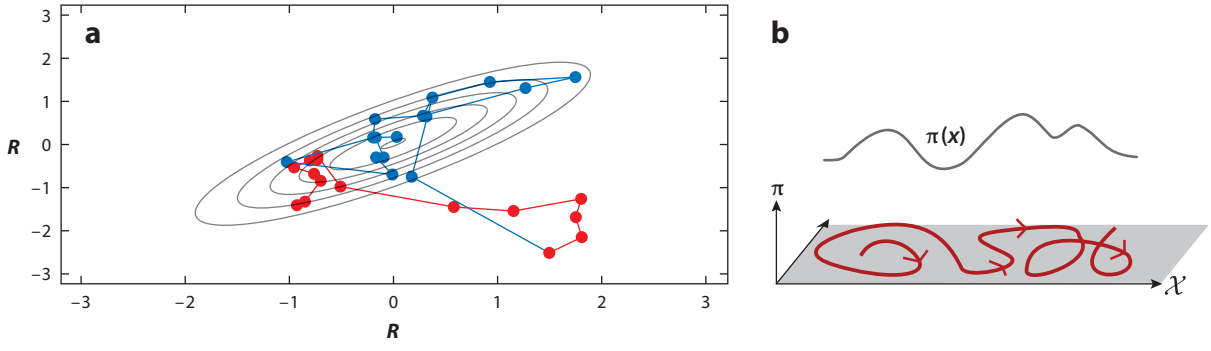

Figure 1

(*a*) Comparison of random-walk Metropolis (*red*) against hybrid Monte Carlo (*blue*). Thirty samples from a peaked Gaussian distribution were plotted for each method. The use of geometry clearly benefits Hamiltonian Monte Carlo. (*b*) Motion of a particle (*red*) over our sample space $\mathcal{X}$.

(Ajay et al. 1998, Fredrickson et al. 2002, Fernández-Pendás et al. 2014), physics (Duane et al. 1987, Mehlig et al. 1992, Landau & Binder 2009, Sen & Biswas 2017), and engineering (Cheung & Beck 2009, Bui-Thanh & Girolami 2014, Lan et al. 2016, Beskos et al. 2016). The extent of the use of HMC is also illustrated by the long list of users of the Stan language (Carpenter et al. 2016; see, e.g., **http://mc-stan.org/citations** for a full list of publications referencing this software). The above is, of course, a far-from-exhaustive list, but it helps illustrate the relevance of HMC in modern computational sciences.

1.3. Outline

The remainder of this article reviews the use of Hamiltonian dynamics in the context of MCMC. Previous reviews of this methodology were provided by Neal (2011) and Betancourt (2017), who focused mainly on the intuition and algorithmic aspects behind the basic version of HMC. Our aim here is somewhat different and complementary: We focus on formalizing the geometrical and physical foundations of the method (see Sections 2 and 3). This deeper theoretical understanding has provided insight into the development of many extensions of HMC (Betancourt et al. 2016). These include Riemannian manifold Hamiltonian Monte Carlo (RMHMC) (Girolami & Calderhead 2011), introduced in Section 4, and shadow Hamiltonian Monte Carlo (SHMC) (Izaguirre & Hampton 2004), discussed in Section 5. We conclude this review with an outline of the most recent research directions in Section 6, including stochastic gradient methods and HMC in infinite-dimensional spaces.

2. GEOMETRY AND PHYSICS

In HMC, the sample space $\mathcal{X}$ is viewed as a (possibly high-dimensional) space called a manifold, over which a motion is imposed. The reader should keep in mind the idea of a fluid particle moving on the sample space (here, the manifold—see **Figure 1*b***); the algorithm proposes new states by following the trajectory of this particle for a fixed amount of time. By coupling the choice of Hamiltonian dynamics to the target density, the new proposals will allow us to explore the density more efficiently by reducing the correlation between samples, and hence make MCMC more efficient. This article seeks to explain why this is the case.

In this section, we provide an accessible introduction to notions of geometry that are required to define Hamiltonian mechanics. Our hope is to provide the bare minimum of geometry in order to provide some insight into the behavior of the Markov chains obtained. The avid reader is referred to Arnold (1989) and Frankel (2012) for a more thorough introduction to geometry and physics, and to Amari (1987) and Murray & Rice (1993) for the interplay of geometry and statistics. In particular, some of the concepts presented here also have a role in the study of statistical estimation, shape analysis, probability distributions on manifolds, and point processes (Kass & Vos 1997, Dryden & Mardia 1998, Chiu et al. 2013, Dryden & Kent 2015).

2.1. Manifolds and Differential Forms

Manifolds generalize the notions of smooth curves and surfaces to higher dimensions and are at the core of modern mathematics and physics. Simple examples include planes, spheres, and cylinders, but more abstract examples include parametric families of statistical models. Manifolds arise by noticing that smooth geometrical shapes and physical systems are coordinate-independent concepts, so their definitions should not rely on any particular coordinate system. Coordinate patches (defined below) assign coordinates to subsets of the manifold and allow us to turn geometric

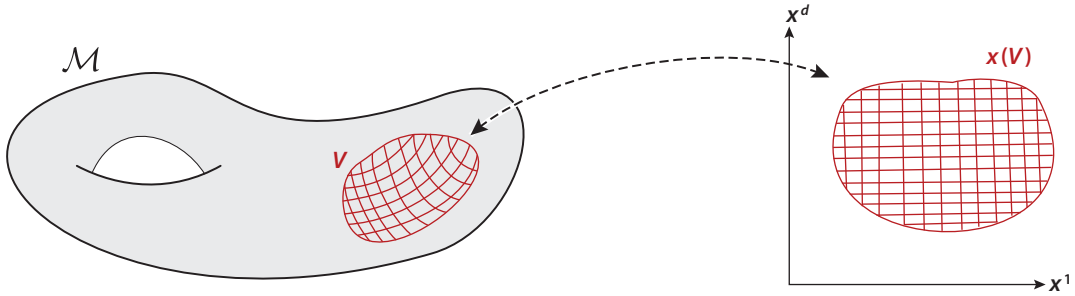

Figure 2

The coordinate patch $\boldsymbol{x}$ attaches coordinates $(x^1, \ldots, x^d)$ to points in the neighborhood $V \subseteq \mathcal{M}$.

questions into algebraic ones. In particular, coordinate patches allow us to transfer the calculus of $\mathbb{R}^d$ to the manifold. It is rarely possible to define a single coordinate patch over the entire manifold, except for the simplest manifolds.

A d-dimensional manifold is a set $\mathcal{M}$ such that every point $q \in \mathcal{M}$ has a neighborhood $V \subseteq \mathcal{M}$ that can be described by d-coordinate functions $(x^1, \ldots, x^d)$.[1] This means that there exists a bijection $\boldsymbol{x}_V : V \to \boldsymbol{x}_V(V) \subseteq \mathbb{R}^d$, called a coordinate patch, which assigns the coordinates $\boldsymbol{x}_V(q) = \big(x^1(q), \ldots, x^d(q)\big)$ to q. The functions $x^j : V \to \mathbb{R}$ are called local coordinates, which we view as being imprinted on the manifold itself (see **Figure 2**). Whenever two patches $\boldsymbol{x}_V, \boldsymbol{x}_W$ overlap, $V \bigcap W \neq \emptyset$, any point q in the overlap is assigned two coordinates $\boldsymbol{x}_W(q), \boldsymbol{x}_V(q)$; in this case, we require the patches to be compatible, i.e., the map $\boldsymbol{x}_V \circ \boldsymbol{x}_W^{-1}$, which is just the map that relates the coordinates, should be smooth (C^∞).

For example, two possible patches for the (1-dimensional) semicircle $x^2 + y^2 = 1$, $y > 0$ in a neighborhood of $(0, 1)$ are $\boldsymbol{x}_V((x, y)) = x$, and $\boldsymbol{\theta}_V((x, y)) = \theta$, where θ satisfies $(\cos\theta, \sin\theta) = (x, y)$. The smoothness of $\boldsymbol{x}_V \circ \boldsymbol{\theta}_V^{-1} = x(\theta) = \cos\theta$ and $\boldsymbol{\theta}_V \circ \boldsymbol{x}_V^{-1} = \theta(x) = \cos^{-1} x$ implies the patches are compatible (see **Figure 3*a***). The sphere S^2 is a 2-(sub)manifold in $\mathbb{R}^3$. In a neighborhood of the north pole, points are specified by their x, y coordinates, since we can write

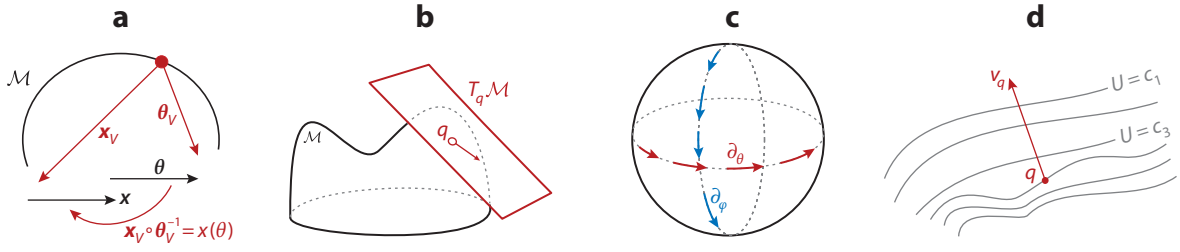

Figure 3

Manifolds and differential forms. (*a*) Patches $\boldsymbol{x}_V, \boldsymbol{\theta}_V$ on the upper hemisphere assign different real numbers to points on the circle $\mathcal{M} = S^1$. (*b*) Tangent vectors at q belong to the tangent space $T_q\mathcal{M}$. (*c*) On a sphere with coordinates (θ, φ), $\theta \in (0, 2\pi)$, $\varphi \in (0, \pi)$, the vector field ∂_θ (*red*) is tangent to the θ-coordinate lines (lines of constant φ). ∂_φ (*blue*) is tangent to the φ-coordinate lines. (*d*) The 1-form $\mathrm{d}_q U$ when applied to v_q tells us how much potential U is gained along the vector v_q.

[1]Technically, for $\mathcal{M}$ to be a manifold, we further require that the topology generated by the differential structure consisting of all compatible patches be Hausdorff and have a countable base; see Arnold (1989) for more details.

z as the graph $z = z(x,y) = \sqrt{1-x^2-y^2}$. These points may be written as $(x, y, z(x,y))$, and we can define a patch $\boldsymbol{x}(x, y, z(x,y)) = (x, y)$. We could have also used the spherical coordinates (θ, φ) on the upper half of S^2.

Leibniz's rule: given a vector v_q at some point $q \in \mathcal{M}$, $v_q(fh) = f(q)v_q(h) + h(q)v_q(f)$

A more interesting example is the statistical manifold of Gaussian distributions $\mathcal{M} = \{\mathcal{N}(\cdot \mid \mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\}$, which is a manifold endowed with global coordinates $\boldsymbol{x}_{\mathcal{M}}\left(\mathcal{N}(\cdot \mid \mu, \sigma^2)\right) = (\mu, \sigma^2)$.

A function $f : \mathcal{M} \to \mathbb{R}$ on the manifold is said to be smooth at a point $q \in \mathcal{M}$ if there exists a coordinate patch $\boldsymbol{x}_V$ around q such that $f_V := f \circ \boldsymbol{x}_V^{-1} : \boldsymbol{x}_V(V) \to \mathbb{R}$ is smooth. The map f_V is just the coordinate expression of f. Since the coordinate patches are compatible, this definition of smoothness is independent of the choice of patches. The space of smooth functions on $\mathcal{M}$ is denoted $C^\infty(\mathcal{M})$.

Example 1. Smooth function on the circle: If $f : S^1 \to \mathbb{R}$, locally around $(0, 1)$, $f \circ \boldsymbol{x}_V^{-1} = f(x)$ and $f \circ \boldsymbol{\theta}_V^{-1} = f(\theta)$.

To define Hamiltonian dynamics, we now introduce the concept of velocity of the flow of a particle on $\mathcal{M}$ (i.e., our sample space) defined by tangent vectors to the manifold. Recall that in $\mathbb{R}^d$, any vector $v = (v^1, \ldots, v^d)$ defines a directional derivative that acts on functions $f \in C^\infty(\mathbb{R}^d)$, by $v(f) := \nabla_v f = v \cdot \nabla f = \sum_{j=1}^d v^j \partial_j f$, where $\partial_i := \frac{\partial}{\partial x^i}$. We can thus think of the vector v as a first-order differential operator $v = \sum_{j=1}^d v^j \partial_j : C^\infty(\mathbb{R}^d) \to \mathbb{R}$ (which is linear and satisfies Leibniz's rule). We now generalize this to manifolds: If $f, h \in C^\infty(\mathcal{M})$, we define a tangent vector $v_q : C^\infty(\mathcal{M}) \to \mathbb{R}$ at $q \in \mathcal{M}$ to be a linear map satisfying Leibniz's rule.

Defining a linear combination of tangent vectors by $(au_q + bv_q)f := au_q(f) + bv_q(f)$ turns the set of tangent vectors at $q \in \mathcal{M}$ into a vector space denoted $T_q\mathcal{M}$, called the tangent space at q (see **Figure 3*b***). Consider a local coordinate patch (V, ϕ_V) around q. The coordinate functions x^j define tangent vectors $\partial_j|_q$ at q by

$$\partial_j|_q(f) := \left.\frac{\partial f_V}{\partial x^j}\right|_{\boldsymbol{x}(q)}.$$

These tangent vectors form a basis of $T_q\mathcal{M}$; any tangent vector at a point q is of the form $\sum_{j=1}^d v^j \partial_j|_q$, where $v^j \in \mathbb{R}$. A vector field v is a smooth map that assigns, at each point q, a tangent vector v_q. Locally any vector field can be written as $v = \sum_{j=1}^d v^j(\boldsymbol{x})\partial_j|_x$, where ∂_j is the (local) vector field $\partial_j : q \mapsto \partial_j|_q$. See **Figure 3*c*** for an example on the sphere.

The objects $\mathrm{d}f$ and $\mathrm{d}x$ are often introduced as being mysterious infinitesimal vectors/quantities that give a real number when integrated. These objects are, in fact, special cases of differential forms, which we now formally introduce; they play a central role in Hamiltonian mechanics.

A 1-form at a point $q \in \mathcal{M}$ is a linear functional on the tangent space, i.e., a linear map $\alpha_q : T_q\mathcal{M} \to \mathbb{R}$. The simplest example is the differential of a function, $\mathrm{d}_q f$, which maps a vector v_q to the rate of change of f in direction v_q: $\mathrm{d}_q f(v_q) := v_q(f)$. In a coordinate patch, we can consider the differential of the coordinate function x^i. Taking $v_q = \partial_j|_q$, we see that $\mathrm{d}_q x^i(\partial_j|_q) = \partial_j|_q(x^i) = \delta^i_j$, where δ^i_j is 1 if $i = j$ and 0 otherwise.

Example 2. Example of a differential: Let (θ, z) be coordinates on a cylinder. Suppose $f_V(\theta, z) := z^2 - \theta$, then $\mathrm{d}f = 2z\mathrm{d}z - \mathrm{d}\theta$. At $q = (1, 3)$, $\mathrm{d}_q f = 6\mathrm{d}_q z - \mathrm{d}_q\theta$. The rate of change of f along $v_q = -2\partial_z|_q$ is $\mathrm{d}_q f(v_q) = -12$.

Hence, $(\mathrm{d}_q x^j)$ is the dual basis to $(\partial_j|_q)$ and a basis of $T^*_q\mathcal{M}$, the vector space of 1-forms at q. A 1-form α is a smooth map that assigns at each point q a 1-form α_q. Locally (i.e., in a given

coordinate patch), any differential 1-form may be written as $\alpha = \sum_{j=1}^{d} \alpha_j(\boldsymbol{x})\mathrm{d}x^j$, where $\mathrm{d}x^j$ is the (local) differential 1-form $\mathrm{d}x^j : q \mapsto \mathrm{d}_q x^j$. For example, the differential of the function f is $\mathrm{d}f = \frac{\partial f_V}{\partial x^1}\mathrm{d}x^1 + \cdots + \frac{\partial f_V}{\partial x^d}\mathrm{d}x^d$.

A physical example of a 1-form is the force F acting on a particle, which is given by the differential of a potential energy function $F = -\mathrm{d}U$. In HMC, the potential energy U is related to the target unnormalized density by $U := -\log(\tilde{\pi})$. Given a vector, the force measures the rate at which potential energy is gained by moving in that direction (see **Figure 3*d***). Directions of increasing U correspond to directions of decreasing probability.

Finally, to define the notions of volume, curvature, and of length of curves on a manifold, it suffices to define the length of tangent vectors. A Riemannian metric[2] g is a smooth assignment of an inner product $g_q : T_q\mathcal{M} \times T_q\mathcal{M} \to \mathbb{R}$ at each point $q \in \mathcal{M}$. The pair $(\mathcal{M}, g)$ is called a Riemannian manifold. Sub-manifolds of $\mathbb{R}^d$ have a natural Riemannian metric which arises by simply restricting the standard inner product of $\mathbb{R}^d$ to the submanifold. In local coordinates, we can define at each point q a symmetric matrix $\boldsymbol{G}$ such that $\boldsymbol{G}_{ij} := g_q(\partial_i|_q, \partial_j|_q)$. We then recover the usual inner product space result $g_q(v, u) = \boldsymbol{v}^T \boldsymbol{G}(\boldsymbol{x}(q))\boldsymbol{u}$, where $\boldsymbol{u}$ is the array $(u^1, \ldots, u^d)$ of coefficients of the vector u in the local coordinate basis $u = u^1\partial_1 + \cdots + u^d\partial_d$.

The tools developed above allow us to formalize Hamiltonian dynamics on manifolds, which will be used to create efficient proposals for MCMC.

Length of curves: the inner product defines a norm $||v||^2 = g(v, v)$; the length of a curve γ is given by integrating its tangent vector $\int_\gamma ||\dot{\gamma}||$

2.2. Hamiltonian Mechanics

Consider a particle moving on $\mathcal{M}$ from initial position $q \in \mathcal{M}$. We call $\mathcal{M}$ the configuration manifold (or configuration space). The particle could, for example, be a mass attached at the end of a plane pendulum (so $\mathcal{M} = S^1$) or a fluid particle flowing along a river. The deterministic motion followed by the particle is governed by the laws of physics. Let $\Phi_t(q)$ be its position at time t, so $\Phi_0(q) = q$, and the trajectory followed by the particle is given by the curve $\gamma : t \mapsto \Phi_t(q)$. The curve γ generates a vector field $\dot{\gamma}$ over the range of γ representing the velocity of the particle; the tangent vector at the point $\gamma(a) = r$ is defined, for any function f, by

$$\dot{\gamma}_r(f) := \left.\frac{df(\gamma(t))}{dt}\right|_{t=a} = (f \circ \gamma)'(a).$$

Since the laws of physics are the same at all times, $\Phi_t \circ \Phi_s(q) = \Phi_{t+s}(q)$. We call Φ the flow and $\dot{\gamma}$ the velocity field (see **Figure 4*a***).

The particle has a kinetic energy K that measures the energy carried by its speed and mass. If no forces are acting, the particle's kinetic energy (and speed) will be constant; otherwise, the force F will increase/decrease the particle's kinetic energy. Since energy is conserved, the particle must be losing/gaining some other type of energy introduced by the force field, which we call potential energy U (see **Figure 4*b*** for an example on the pendulum). It can be shown that $F = -\mathrm{d}U$, so the force is caused by variations in potential energy.

A Riemannian metric provides an identification between vector fields and differential 1-forms by associating the vector field v to the 1-form $\alpha(\cdot) := g(v, \cdot)$, and the inner product on vectors $g(u, v) = \boldsymbol{u}^T \boldsymbol{G}v$ induces an inner product on the associated 1-forms (if and only if $\det \boldsymbol{G} \neq 0$) by $g^{-1}(p, \alpha) := \boldsymbol{p}^T \boldsymbol{G}^{-1}\boldsymbol{\alpha}$. In particular each velocity field $\dot{\gamma}$ induced by a curve γ has an associated momentum field defined by $p := g(\dot{\gamma}, \cdot)$ which represents the "quantity of motion" in direction $\dot{\gamma}$.

[2] Riemannian geometry was introduced in statistics by Rao, who noted the Fisher-Rao metric defined a useful notion of distance between populations.

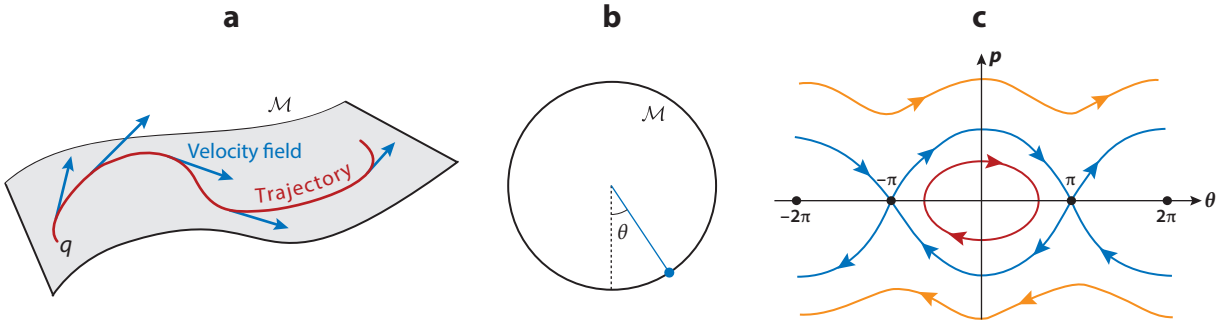

Figure 4

(*a*) The particle initially at q follows the trajectory $\gamma = \Phi(q)$ (*red*). Its velocity field $\dot{\gamma}$ is in blue. (*b*) A mass (*blue*) attached to a pendulum has $\mathcal{M} = S^1$. As the mass moves from $\theta = 0$ to $\theta = \pi$, its kinetic energy is transformed into potential energy. (*c*) Possible trajectories in phase space $\mathcal{N}$ for the mass pendulum. The red/blue/orange trajectories represent respectively the cases when there is not enough/exactly enough/more than enough energy to do a full turn.

Writing $p = \sum_{j=1}^{d} p_j(\boldsymbol{x})\mathrm{d}x^j$ makes it clear that to define p (i.e., to specify each 1-form p_q, called momentum, defined by p at $q \in \mathcal{M}$) we need to specify the $2d$-tuple: $(\boldsymbol{x}(q), \boldsymbol{p}(q)) := \big(x^1(q), \ldots, x^d(q), p_1(q), \ldots, p_d(q)\big)$, i.e., the position $\boldsymbol{x}(q)$ of p_q and the momentum components at $\boldsymbol{p}(q)$.[3]

> **Example 3.** Example of a momentum field: Suppose a particle in a plane has momentum field $p = ye^x\mathrm{d}x - x\mathrm{d}y$. Then $\boldsymbol{z} = (x, y, ye^x, -x)$. At $q = (1, 3)$ its momentum is $p_q = 3ed_q x - d_q y$ and its phase is $(1, 3, 3e, -1)$.

Thus, the set of momenta (or equivalently the set of 1-forms) $T^*\mathcal{M} = \bigcup_q T^*_q\mathcal{M}$ is a $2d$-dimensional manifold, called the cotangent bundle, on which $\boldsymbol{z} := (\boldsymbol{x}, \boldsymbol{p}) = (x^1, \ldots, x^d, p_1, \ldots, p_d)$ are coordinates.

At any given time t, the $2d$-tuple $\boldsymbol{z}(\gamma(t))$, consisting of the position $\boldsymbol{x}(\gamma(t))$ of the particle and its momentum $\boldsymbol{p}(\gamma(t))$, is called the phase and fully specifies the physical system, i.e., it encodes all the information about the system and determines its future dynamics. The space of all possible phases is called phase space or the cotangent bundle $T^*\mathcal{M}$ (see **Figure 4*c***) for the phase space of the pendulum example).

Bundle: locally a Cartesian product of manifolds, but globally may be twisted like a Möbius strip

Symplectic 2-form: the symplectic 2-form ω turns $\mathrm{d}H$ into a trajectory γ through $\omega(\dot{\gamma}, \cdot) = \mathrm{d}H$; the properties of ω ensure γ is compatible with physics

Forces acting on the system may be accounted by defining how the energy transfers between potential and kinetic (since $F = -\mathrm{d}U$). Hence, if we define the Hamiltonian function H to be the total energy $H = K + U$, we expect its differential $\mathrm{d}H$ to fully determine the dynamics of the system (from now on, K is viewed as a function of the momentum rather than the velocity); see **Figure 5*a***. We now construct Hamiltonian mechanics, in which the trajectory of the particle on $\mathcal{M}$ is described by a trajectory in phase space $\mathcal{N} := T^*\mathcal{M}$ defining how the phase of the system evolves. From here on, Φ is a flow on $\mathcal{N}$ and γ a curve $t \mapsto \Phi_t(\boldsymbol{z}_0)$ on $\mathcal{N}$ for some initial phase $\boldsymbol{z}_0$ [locally, $\gamma(t)$ is now described by coordinates $\boldsymbol{z}(t) := (\boldsymbol{x}(\gamma(t)), \boldsymbol{p}(\gamma(t)))$ and $\boldsymbol{x}(\gamma(t))$ are the coordinates of the physical trajectory in $\mathcal{M}$]. To do so we will need a map that turns $\mathrm{d}H$ into a trajectory γ that is consistent with the laws of physics. This map is called a symplectic 2-form, and we now describe it.

[3] In general, $p := \partial\mathcal{L}/\partial v$ where $\mathcal{L}(x, v)$ is the Lagrangian (see the **Supplemental Appendix**). If $\mathcal{L} = K - U$ with $K = g(v, v)/2 = g(\dot{\gamma}, \dot{\gamma})/2$ and if g is constant then $p = g(\dot{\gamma}, \cdot)$; but this is not true in general.

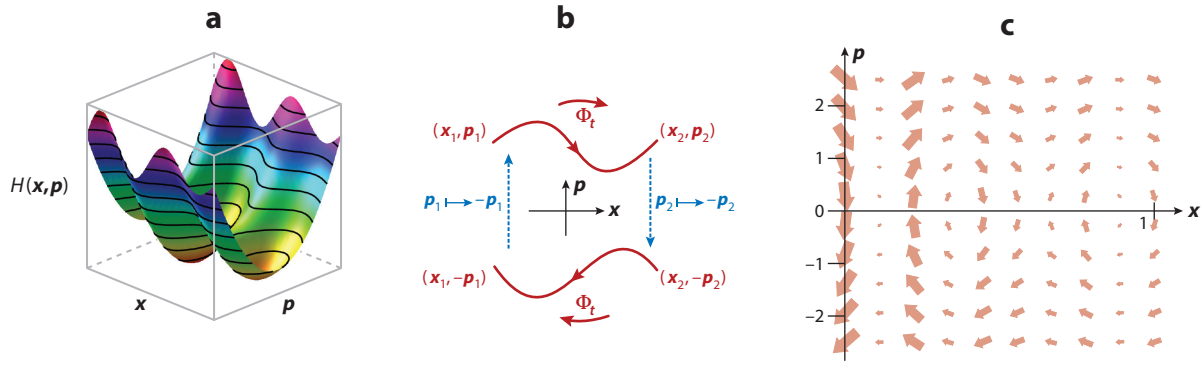

Figure 5

(*a*) Surface plot of a Hamiltonian with contour lines. (*b*) Time-reversibility of Hamiltonian mechanics. (*c*) Hamiltonian vector field for the system shown in panel *a*.

We need at each phase $\mathbf{z} \in \mathcal{N}$ an invertible linear map (since Newton's equations are linear) $S_{\mathbf{z}}^{-1} : T_{\mathbf{z}}^*\mathcal{N} \to T_{\mathbf{z}}\mathcal{N}$ to turn the differential form $\mathrm{d}H$ into the vector field $\dot{\gamma}$ generated by the trajectory in phase space γ (this vector field yields the velocity field when projected to the configuration space $\mathcal{M}$). Its inverse, $S_{\mathbf{z}} : T_{\mathbf{z}}\mathcal{N} \to T_{\mathbf{z}}^*\mathcal{N}$, maps linearly vectors into 1-forms and fully determines Hamiltonian dynamics. Any such linear map $S_{\mathbf{z}}$ may be identified with a bilinear map $\omega_{\mathbf{z}} : T_{\mathbf{z}}\mathcal{N} \times T_{\mathbf{z}}\mathcal{N} \to \mathbb{R}$ where $\omega_{\mathbf{z}}(u, v) = (S_{\mathbf{z}}(u))(v)$. Letting ω be the smooth map $\mathbf{z} \mapsto \omega_{\mathbf{z}}$, note that since $S(\dot{\gamma}) = \mathrm{d}H$, then $\omega(\dot{\gamma}, \cdot) = \mathrm{d}H(\cdot)$, i.e., $\omega(\dot{\gamma}, \cdot)$ maps a vector field to the rate of change of H along it. A differential 2-form β is a smooth map that assigns to each $\mathbf{z} \in \mathcal{N}$ a bilinear, antisymmetric map $\beta_{\mathbf{z}} : T_{\mathbf{z}}\mathcal{N} \times T_{\mathbf{z}}\mathcal{N} \to \mathbb{R}$. We now show that ω is a symplectic 2-form (also called a symplectic structure), i.e., it satisfies the following:

1. Nondegenerate differential 2-form: By the law of conservation of energy, the total energy of the system must be constant, $\frac{\mathrm{d}}{\mathrm{d}t}(H \circ \gamma(t)) = 0$, or equivalently, $\mathrm{d}H(\dot{\gamma}) = 0$. Thus, for all flows, we have $\omega(\dot{\gamma}, \dot{\gamma}) = 0$, which implies that ω is antisymmetric and thus a differential 2-form. Moreover ω is nondegenerate, which means the velocity field $\dot{\gamma}$ exists globally.
2. Closed: The laws of physics must be conserved in time, which means that ω is conserved along the flow and is ensured by demanding that its differential vanishes, i.e., $\mathrm{d}\omega = 0$ (the differential of a 2-form is formally defined in the **Supplemental Appendix**). This gives rise to conservation of volume: If particles are initially occupying a region U in phase space with volume $\mathrm{vol}(U)$, this volume will be preserved as they follow the flow, i.e., $\mathrm{vol}(\Phi_t(U)) = \mathrm{vol}(U)$.

Supplemental Material

When $\mathcal{M} = \mathbb{R}^d$, the phase space $\mathcal{N} = \mathbb{R}^{2d}$ has a natural symplectic structure, which in (global) coordinates $(x^1, \ldots, x^d, p_1, \ldots, p_d)$ is given by

$$\omega := \mathrm{d}x^1 \wedge \mathrm{d}p_1 + \cdots + \mathrm{d}x^d \wedge \mathrm{d}p_d.$$

Here, $\mathrm{d}_{\mathbf{z}}x^i \wedge \mathrm{d}_{\mathbf{z}}p_j : T_{\mathbf{z}}\mathcal{N} \times T_{\mathbf{z}}\mathcal{N} \to \mathbb{R}$ is the 2-form constructed using the wedge product that, given a pair of vectors, gives the signed area of the parallelogram spanned by their projection to the x^i–p_j plane (see the **Supplemental Appendix**).

Bilinear map: a map that is linear in each of its arguments

Example 4. Example of wedge product: In $\mathbb{R}^3$, $\mathrm{d}y \wedge \mathrm{d}z$ applied to $(3, 2, -5)$ and $(1, 7, 4)$ gives the signed area of parallelogram spanned by $(2, -5)$ and $(7, 4)$.

The condition $\omega(\dot{\gamma}, \cdot) = \mathrm{d}H(\cdot)$ implies that the coordinate expression of γ, $(x^1(t), \ldots, x^d(t), p_1(t), \ldots, p_d(t))$, satisfies Hamilton's equations,

$$\frac{\mathrm{d}x^i}{\mathrm{d}t} = \frac{\partial H}{\partial p_i}, \qquad \frac{\mathrm{d}p_i}{\mathrm{d}t} = -\frac{\partial H}{\partial x^i}, \tag{1.}$$

i.e., the velocity field is orthogonal to the gradient of H. These can be rewritten as a single equation,

$$\frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}t} = J\frac{\partial H}{\partial \boldsymbol{z}},$$

where J is the canonical symplectic matrix given by

$$J = \begin{pmatrix} \mathbf{0} & I_{d\times d} \\ -I_{d\times d} & \mathbf{0} \end{pmatrix}.$$

Finally, notice Hamiltonian mechanics is time reversible, i.e., Equation 1 is preserved under the transformation $t \to -t$, $\boldsymbol{x} \to \boldsymbol{x}$, $\boldsymbol{p} \to -\boldsymbol{p}$. This means the following: Consider a system, say a pendulum, with initial state $(\boldsymbol{x}_1, \boldsymbol{p}_1)$. After a time t it will have a state $(\boldsymbol{x}_2, \boldsymbol{p}_2) = \Phi_t(\boldsymbol{x}_1, \boldsymbol{p}_1)$. If we reverse its momentum, $(\boldsymbol{x}_2, \boldsymbol{p}_2) \mapsto (\boldsymbol{x}_2, -\boldsymbol{p}_2)$, then after another time t it will be at its initial position with opposite momentum, i.e., $\Phi_t(\boldsymbol{x}_2, -\boldsymbol{p}_2) = (\boldsymbol{x}_1, -\boldsymbol{p}_1)$ (see **Figure 5*b***). Time-reversibility is necessary for detailed balance to hold in MCMC.

We have now defined the basic notions necessary to define Hamiltonian dynamics. More precisely, we have explained how the motion of a fluid particle on the manifold $\mathcal{M}$ is described by a curve in phase space $\mathcal{N} = T^*\mathcal{M}$. For this curve to represent a physical path, we have shown it must be related to the differential of the Hamiltonian $\mathrm{d}H$ through a symplectic form ω.

3. HAMILTONIAN MONTE CARLO

3.1. Hamiltonian Dynamics

In practice, Hamilton's equations (see the sidebar Hamiltonian Mechanics) cannot be solved exactly, and we must employ numerical methods that approximate the flow in Equation 1 (Leimkuhler & Reich 2004, Hairer et al. 2006). Let $\boldsymbol{z}(t_0) := (\boldsymbol{x}(t_0), \boldsymbol{p}(t_0))$ be the initial phase of a Hamiltonian system H.

If we fix a time-step τ, we can obtain a sequence of points along the trajectory that describe how the phase evolves:

$$\boldsymbol{z}(t_0) \to \boldsymbol{z}(t_1) := \Phi_\tau(\boldsymbol{z}(t_0)) \to \boldsymbol{z}(t_2) := \Phi_\tau(\boldsymbol{z}(t_1)) = \Phi_\tau^2(\boldsymbol{z}(t_0)) \to \cdots$$

A numerical one-step method is a map Ψ_τ that approximates this trajectory

$$\boldsymbol{z}(t_0) \to \boldsymbol{z}^1 := \Psi_\tau(\boldsymbol{z}(t_0)) \to \boldsymbol{z}^2 := \Psi_\tau(\boldsymbol{z}^1) = \Psi_\tau^2(\boldsymbol{z}(t_0)) \to \ldots$$

HAMILTONIAN MECHANICS

William Rowan Hamilton developed Hamiltonian mechanics as a generalization of classical dynamics by applying ideas from optics and by reformulating Lagrangian mechanics. A more general introduction to Hamiltonian and Lagrangian dynamics presented in the **Supplemental Appendix** may be of interest to readers interested in gaining a deeper understanding of some of the more advanced Hamiltonian Monte Carlo methods.

Supplemental Material

where $z^k = (x^k, p^k)$ approximates $z(t_k)$. The numerical method will introduce an error at each step, defined as the difference between the application of Φ_τ and Ψ_τ to a phase z.

Such errors will accumulate over time, and the approximated trajectory will gradually deviate from the exact one. To partially remedy this, we make use of geometric integrators, which are numerical methods that exactly preserve some fundamental properties of the dynamics they simulate, and hence ensure that the approximated trajectory retains some key features. In particular, symplectic integrators are geometric integrators that preserve the symplectic structure ω and thus the volume in phase space.

Symplectic map: symplectic map $S : \mathcal{M} \to \mathcal{M}$ is one for which the induced map on 2-forms preserves ω, $S_* : \omega \mapsto \omega$

Any smooth map $S : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ has a Jacobian matrix S_z at each phase $z = (x, p)$, which is a linear map $T_z\mathbb{R}^{2d} \to T_z\mathbb{R}^{2d}$. We say S is a symplectic map if $S_z^T J^{-1} S_z = J^{-1}$, where J is the canonical symplectic matrix. The method Ψ_τ is called a symplectic integrator if it is a symplectic map. Writing $z^{k+1} = \Psi_\tau(z^k)$, this is equivalent to requiring that it preserves the symplectic structure for each step k:

$$\mathrm{d}x^{k+1} \wedge \mathrm{d}p^{k+1} = \mathrm{d}x^k \wedge \mathrm{d}p^k.$$

A useful technique to build symplectic integrators uses Hamiltonian splitting.

Suppose our Hamiltonian is of the form $H = H_1 + \cdots + H_\ell$, where Hamilton's equations may be solved explicitly for each Hamiltonian H_i. If we denote by $\Phi_\tau^{H_k}$ the exact flow of H_k, we can define a numerical method for H by

$$\Psi_\tau := \Phi_\tau^{H_1} \circ \cdots \circ \Phi_\tau^{H_\ell}.$$

The composition of these exact flows may not give the exact flow of H. However, since each flow $\Phi_\tau^{H_k}$ is symplectic, and the composition of symplectic maps is symplectic, Ψ_τ will be a symplectic integrator. The most popular symplectic integrator is the Störmer–Verlet or leapfrog integrator (see **Figure 6*a***), which is derived through the splitting (see **Figure 6*b***) $H_1 = \frac{1}{2}U(x)$, $H_2 = K(p)$ and $H_3 = \frac{1}{2}U(x)$, which gives

$$\begin{aligned} p^{k+\frac{1}{2}} &= p^k - \frac{\tau}{2}\frac{\partial U}{\partial x}\left(x^k\right), \\ x^{k+1} &= x^k + \tau\frac{\partial K}{\partial p}\left(p^{k+\frac{1}{2}}\right), \text{ and} \\ p^{k+1} &= p^{k+\frac{1}{2}} - \frac{\tau}{2}\frac{\partial U}{\partial x}\left(x^{k+1}\right). \end{aligned}$$

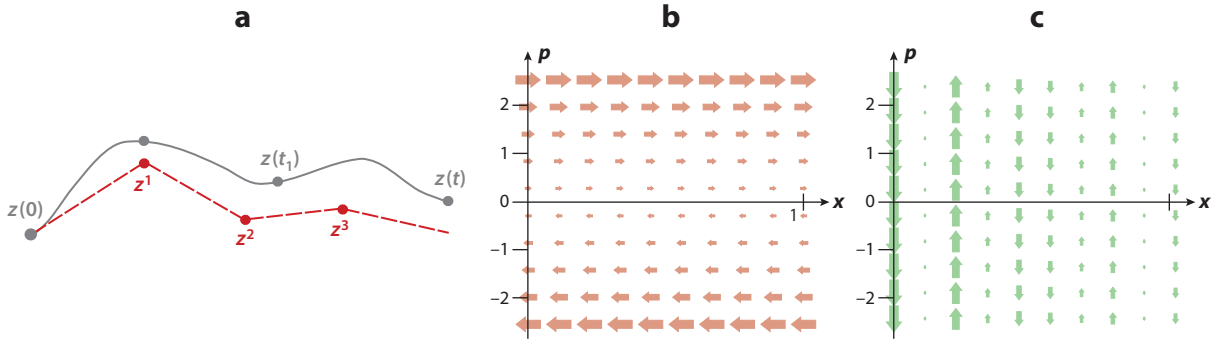

Figure 6

(*a*) Exact evolution of the phase $z(t)$ (*gray curve*) and numerical one-step method (*red dashed curve*). (*b–c*) Hamiltonian vector fields for K and U of the system in **Figure 5*a***, respectively, both of which can be integrated exactly.

It is easy to verify that the leapfrog integrator is reversible, i.e., we can invert the leapfrog trajectory by simply negating the momentum, applying the leapfrog algorithm, and negating the momentum again. It is also symmetric: $\Psi_{-\tau}^{-1} = \Psi_{\tau}$. Reversibility and conservation of volume of the integrator are required to prove detailed balance when we apply it in HMC. However, the energy is only approximately conserved along a leapfrog trajectory.

The leapfrog integrator is an integrator of order 2, which means that its global error is of order τ^3, where τ is the step size. In situations in which very high accuracy is needed, it may be necessary to turn to higher-order integrators to obtain better approximations of the exact trajectory over a short time interval (Campostrini & Rossi 1990, Yoshida 1990, Leimkuhler & Reich 2004). The improved accuracy must, however, be balanced with the increased computational cost. Other integrators have also been proposed; see, for example, Blanes et al. (2014).

3.2. The Hamiltonian Monte Carlo Algorithm

Suppose we want to sample from a probability density $\pi_{\boldsymbol{x}} : \mathcal{X} \to \mathbb{R}$, which we only know up to multiplicative constant: $\pi_{\boldsymbol{x}} = \tilde{\pi}_{\boldsymbol{x}}/Z$. The differential of $U(\boldsymbol{x}) := -\log \pi_{\boldsymbol{x}}(\boldsymbol{x})$, if it is known, informs us what directions lead to regions of higher probability. It can also be computed without knowledge of Z. In HMC, we view $U(\boldsymbol{x})$ as being a potential energy (Duane et al. 1987), which enables us to rewrite the target density as

$$\pi_{\boldsymbol{x}}(\boldsymbol{x}) = \frac{1}{Z} \exp\left(-U(\boldsymbol{x})\right).$$

We then interpret regions of higher potential energy as regions of lower probability. The state space $\mathcal{X}$ plays the role of the configuration manifold $\mathcal{M}$ on which the dynamics are defined. We define Hamiltonian dynamics on $\mathcal{X}$ by introducing a kinetic energy $K(\boldsymbol{p}) = \frac{1}{2} g^{-1}(\boldsymbol{p}, \boldsymbol{p}) = \frac{1}{2} \boldsymbol{p}^T \boldsymbol{G}^{-1} \boldsymbol{p}$, and thus a Hamiltonian $H(\boldsymbol{x}, \boldsymbol{p}) = K + U$. We view the $d \times d$ matrix $\boldsymbol{G}$ as a covariance matrix and assume that the momentum variables have the multivariate Gaussian density

$$\pi_{\boldsymbol{p}}(\boldsymbol{p}) = \mathcal{N}(\boldsymbol{p}; \boldsymbol{0}, \boldsymbol{G}) = \left((2\pi)^d |\boldsymbol{G}|\right)^{-\frac{1}{2}} \exp\left(-K(\boldsymbol{p})\right),$$

where $|\boldsymbol{G}|$ denotes the determinant of $\boldsymbol{G}$. The choice of the matrix $\boldsymbol{G}$ is critical for the performance of the algorithm, yet there is no general principle guiding its tuning, so it is often set to be the identity matrix. In Section 4, we will see how the local structure of the target density may be used to choose a position-dependent $\boldsymbol{G}$. Define a joint density by

$$\pi(\boldsymbol{x}, \boldsymbol{p}) = Z^{-1} \left((2\pi)^d \mid \boldsymbol{G} \mid\right)^{-\frac{1}{2}} \exp\left(-H(\boldsymbol{x}, \boldsymbol{p})\right) = \pi_{\boldsymbol{x}}(\boldsymbol{x}) \pi_{\boldsymbol{p}}(\boldsymbol{p}).$$

The HMC algorithm (see the sidebar Hamiltonian Monte Carlo Steps) generates samples from this joint density. Since the total energy H is preserved along the flow, the joint probability $\pi(\boldsymbol{x}, \boldsymbol{p})$ is constant along Hamiltonian trajectories. Here the Hamiltonian splitting $H = K + U$ is clearly

HAMILTONIAN MONTE CARLO STEPS

Step 1 of the algorithm is a momentum heat bath (Gibbs sampler). Step 2 is a molecular dynamics step (2a) followed by a Markov chain Monte Carlo (MC) rejection step (2b). This is sometimes called the Metropolis–Hastings step, although neither of them had much to do with it!

applicable and we can hence use the leapfrog integrator. In practice, the simulation will not be exact since the leapfrog integrator is only approximately energy-preserving, and a Metropolis step will be necessary to ensure that we sample from the correct joint density. Given a current phase $(\boldsymbol{x}^k, \boldsymbol{p}^k) \in T^*\mathcal{X}$, the algorithm at iteration k is:

1. Draw a momentum variable $\boldsymbol{p}^{k'}$ using $\pi_{\boldsymbol{p}}(\boldsymbol{p})$ i.e., $\boldsymbol{p}^{k'} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{G})$.

2a. Simulate dynamics with initial phase $(\boldsymbol{x}^k, \boldsymbol{p}^{k'})$ using the leapfrog integrator with fixed step-size τ for L leapfrog steps, and flip (i.e., negate) the momentum of the resulting phase. This yields a proposal phase $(\boldsymbol{x}^*, \boldsymbol{p}^*)$.

2b. Accept the phase $(\boldsymbol{x}^*, \boldsymbol{p}^*)$ using a Metropolis step with probability

$$\min[1, \exp(-H(\boldsymbol{x}^*, \boldsymbol{p}^*) + H(\boldsymbol{x}^k, \boldsymbol{p}^{k'}))],$$

else keep the current phase: $(\boldsymbol{x}^{k+1}, \boldsymbol{p}^{k+1}) = (\boldsymbol{x}^k, \boldsymbol{p}^{k'})$.

This algorithm simulates a Markov chain which, if ergodic, converges to the unique stationary density $\pi(\boldsymbol{x}, \boldsymbol{p})$. The Markov chain can be shown to be geometrically ergodic under regularity assumptions (Livingstone et al. 2015). As $\pi_{\boldsymbol{x}}(\boldsymbol{x})$ is a marginal density of our target density $\pi(\boldsymbol{x}, \boldsymbol{p})$, we can simply discard the auxiliary momentum samples to obtain samples of $\pi_{\boldsymbol{x}}(\boldsymbol{x})$.

Two parameters need to be tuned in order to apply HMC: the time-step τ and the trajectory length L. This tuning is often performed by running a few preliminary runs. On the one hand, small time-steps will waste computational resources and slow down the exploration of the sample space, while large values of τ can lead to bad approximations of the trajectory that dramatically reduce the acceptance probability. On the other hand, L needs to be large enough to permit efficient explorations that avoid random walks and generate distant proposals; however, long trajectories may contain points in which the momentum sign flips, which can lead to poor exploration (think of a pendulum) (Neal 2011). Several approaches to tuning have been proposed, the most popular of which appears in Beskos et al. (2013), which proposes to tune parameters to maximize the computational efficiency as $d \to \infty$. Other approaches include the no U-turn sampler (NUTS) algorithm (Hoffman & Gelman 2014), currently in use in the Stan programming software, and the use of Bayesian optimization (Wang et al. 2013). Finally, the shadow HMC algorithm, introduced in Section 5, has also been used to this effect (Kennedy et al. 2012).

3.3. Relations to Stochastic Differential Equations

More information about the dynamics can be preserved (thus making the trajectory more physical) if the full momentum resampling (the first step of HMC) is replaced by a partial momentum replacement (Horowitz 1991, Campos & Sanz-Serna 2015). This enables us to sample more often, as the trajectory length may be reduced to a single time-step without performing a random walk. Let $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{G})$ be Gaussian noise. The generalized HMC algorithm is given by the following steps at each iteration k:

1. Rotate $(\boldsymbol{p}^k, \boldsymbol{\xi}^k)$ by an angle ϕ.

2a. Perform step 2 of the HMC algorithm to reach phase $(\boldsymbol{x}^*, \boldsymbol{p}^*)$.

2b. Flip the momentum, $F : \boldsymbol{p}^* \mapsto -\boldsymbol{p}^*$.

2c. Apply a Metropolis accept/reject step.

3. Flip the momentum, $F : \boldsymbol{p}^{k+1} \mapsto -\boldsymbol{p}^{k+1}$.

When $\phi = \pi/2$, we recover HMC. The first momentum flip is required to satisfy detailed balance, but it means that momentum is reversed in case of rejection, which slows down the exploration if the rejection probability is nonnegligible.

We now briefly mention links between HMC and algorithms based on stochastic differential equations (SDEs). If we consider the HMC algorithm in the special case of a single step of leapfrog integrator (i.e., $L = 1$) with $K(\boldsymbol{p}) = \frac{1}{2}\boldsymbol{p}^T\boldsymbol{p}$ and drop the acceptance step, then each iteration k is equivalent to

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \tau\boldsymbol{p}^{k+1} - \frac{1}{2}\frac{\partial U}{\partial \boldsymbol{x}}\tau^2.$$

Christoffel symbols: give information about the curvature of the manifold and are defined by $\Gamma^k_{ij} = \sum_r \frac{1}{2}g^{kr}(\partial_j g_{ir} + \partial_i g_{rj} - \partial_r g_{ij})$

Defining ε to be the square of the step size τ and the initial momentum to be Gaussian noise $\boldsymbol{\xi}$, we end up with a discretization of the overdamped Langevin equation: $\boldsymbol{x}(\sqrt{\varepsilon}) = \boldsymbol{x}(0) + \sqrt{\varepsilon}\boldsymbol{\xi} - \frac{1}{2}\frac{\partial U}{\partial \boldsymbol{x}}\varepsilon$. If we add a Metropolis–Hastings step, this algorithm corresponds to the MALA previously discussed, which is an exact version of the Langevin algorithm (in the sense that there is no discretization error).

HMC can also be related to higher-order SDEs: Consider the following second-order Langevin dynamics defined on a Riemannian manifold (with diffusion defined by a vector field v),

$$\mathrm{d}\boldsymbol{x} = \boldsymbol{v}\mathrm{d}t, \qquad \mathrm{d}\boldsymbol{v} = -\gamma(\boldsymbol{x}, \boldsymbol{v})\mathrm{d}t - \boldsymbol{G}^{-1}(\boldsymbol{x})\frac{\partial U}{\partial \boldsymbol{x}}\mathrm{d}t - \boldsymbol{v}\mathrm{d}t + \sqrt{2\boldsymbol{G}^{-1}(\boldsymbol{x})}\mathrm{d}W.$$

Here, W is a standard Wiener process and $\mathrm{d}\boldsymbol{v} + \gamma\mathrm{d}t$ is the covariant time derivative (physically, the acceleration) of the velocity and thus γ has kth component $\sum_{ij}\Gamma^k_{ij}v^i v^j$, where Γ^k_{ij} are the Christoffel symbols. The SDE represents the acceleration of a particle on a manifold under the influence of a noisy potential and subject to a friction term $v\,\mathrm{d}t$. The SDE may be transferred to phase space using $\boldsymbol{p} = G(\boldsymbol{x})\boldsymbol{v}$. The invariant distribution of this diffusion may be easily shown to be $\pi(\boldsymbol{x}, \boldsymbol{p}) \propto |G(\boldsymbol{x})|^{-1}\exp(-\frac{1}{2}\boldsymbol{p}^T\boldsymbol{G}^{-1}(\boldsymbol{x})\boldsymbol{p} - U(\boldsymbol{x}))$. Thus, setting $U(\boldsymbol{x}) = -\log\pi_{\boldsymbol{x}}(\boldsymbol{x}) - \frac{1}{2}\log|\boldsymbol{G}(\boldsymbol{x})|$ gives $\pi_{\boldsymbol{x}}(\boldsymbol{x})$ as the marginal distribution.

To simulate from the SDEs above, it is convenient to use schemes that rely on Lie–Trotter splitting (Abdulle et al. 2015), in which the numerical method is an integrator of the form $\Phi_\tau \circ \Psi_\tau$, where Φ_τ is an integrator for the deterministic part and Ψ_τ for the stochastic part. The stochastic part is a conditioned Ornstein–Uhlenbeck process and corresponds to partial momentum refreshment. We can then use a symplectic integrator to sample from $\pi(\boldsymbol{x}, \boldsymbol{p})$ via RMHMC, which we introduce below.

4. RIEMANNIAN MANIFOLD HAMILTONIAN MONTE CARLO

We have seen how HMC uses gradient information from the target density to improve the exploration of the state space. Girolami & Calderhead (2011) introduced the RMHMC method, which uses higher-order information so that the transition density adapts to the local geometry of the target density (see also Livingstone & Girolami 2014). A notion of distance is defined between points in state space, so that smaller steps are performed in directions in which the target density changes rapidly. This method hence borrows tools from information geometry (Amari 1987).

In the original version of this algorithm, Girolami & Calderhead (2011) considered sampling from a Bayesian posterior density: After observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ have been made, the target density $\pi_{\boldsymbol{x}}(\cdot)$ may be updated to a posterior $\pi(\cdot|\boldsymbol{y})$ through a likelihood function $\mathcal{L}(\cdot|\boldsymbol{x})$ by the means of Bayes' theorem, $\pi(\boldsymbol{x}|\boldsymbol{y}) \propto \mathcal{L}(\boldsymbol{y}|\boldsymbol{x})\pi_{\boldsymbol{x}}(\boldsymbol{x})$. They took advantage of the fact that the likelihood function defines a statistical model with parameters $\boldsymbol{x}$, that is, for each $\boldsymbol{x}$, $\mathcal{L}(\cdot|\boldsymbol{x})$ is a density. Under mild conditions (Amari 1987), the statistical model $\mathcal{S} := \{\mathcal{L}(\cdot|\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\}$ is a manifold with global coordinates $\boldsymbol{x}$. Hence, each point on the original manifold $\mathcal{X}$ is now associated to the density $\mathcal{L}(\cdot|\boldsymbol{x})$.

On the statistical manifold $\mathcal{S}$, it is common to identify a vector $v = \sum_{j=1}^d v^j\partial_j$ at $\boldsymbol{x}$ with the random variable (i.e., the function) $v^{(1)}(\cdot) = \sum_{j=1}^d v^j\frac{\partial l(\cdot|\boldsymbol{x})}{\partial x^j}$, where $l(\cdot|\boldsymbol{x}) := \log\mathcal{L}(\cdot|\boldsymbol{x})$ is the

log-likelihood. This is called the 1-representation of the tangent space. We can define a natural inner product on the tangent spaces of $\mathcal{S}$ called the Fisher metric, by defining an inner product on the corresponding 1-representations: $g_{\boldsymbol{x}}(u, v) := \mathbb{E}_{l(\cdot|\boldsymbol{x})}[u^{(1)}v^{(1)}]$. As a result, the configuration manifold $\mathcal{X}$ acquires the Riemannian metric g and thus a natural concept of distance between densities associated to $\boldsymbol{x} \in \mathcal{X}$.

To tailor the metric to Bayesian problems, which are common in MCMC, Girolami & Calderhead (2011) proposed a variant of the Fisher metric, which adds the negative Hessian of the log-prior:

$$\boldsymbol{G}_{ij}(\boldsymbol{x}) \;=\; \boldsymbol{F}_{ij}(\boldsymbol{x}) - \frac{\partial^2}{\partial x^i \partial x^j} \log \pi_0(\boldsymbol{x}),$$

where $\pi_0(\boldsymbol{x})$ is the prior density and $\boldsymbol{F}$ is the Fisher metric. The kinetic energy is defined using this metric $\boldsymbol{G}(\boldsymbol{x})$, so the momentum variable is now Gaussian with a position-dependent covariance matrix $\pi(\boldsymbol{p}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}(\boldsymbol{x}))$, which can help mitigate some of the scaling and tuning issues associated with HMC. The Hamiltonian on the Riemannian manifold is

$$H(\boldsymbol{x}, \boldsymbol{p}) = U(\boldsymbol{x}) + \frac{1}{2} \log\left((2\pi)^d |\boldsymbol{G}(\boldsymbol{x})|\right) + \frac{1}{2} \boldsymbol{p}^T \boldsymbol{G}^{-1}(\boldsymbol{x}) \boldsymbol{p}.$$

The joint density $\pi(\boldsymbol{x}, \boldsymbol{p}) := \exp(-H(\boldsymbol{x}, \boldsymbol{p})) = \pi_{\boldsymbol{x}}(\boldsymbol{x})\pi(\boldsymbol{p}|\boldsymbol{x})$ still has the desired target $\pi_{\boldsymbol{x}}(\boldsymbol{x})$ as marginal density, but the Hamiltonian is no longer separable. Thus, the leapfrog integrator is no longer symplectic and reversible; instead, we use a generalized leapfrog algorithm. At each iteration k, the algorithm is

$$\begin{aligned}
\boldsymbol{p}^{k+\frac{1}{2}} &= \boldsymbol{p}^k - \frac{\tau}{2} \frac{\partial H}{\partial \boldsymbol{x}} \left(\boldsymbol{x}^k, \boldsymbol{p}^{k+\frac{1}{2}}\right), \\
\boldsymbol{x}^{k+1} &= \boldsymbol{x}^k + \frac{\tau}{2} \left(\frac{\partial H}{\partial \boldsymbol{p}} \left(\boldsymbol{x}^k, \boldsymbol{p}^{k+\frac{1}{2}}\right) + \frac{\partial H}{\partial \boldsymbol{p}} \left(\boldsymbol{x}^{k+1}, \boldsymbol{p}^{k+\frac{1}{2}}\right) \right), \\
\boldsymbol{p}^{k+1} &= \boldsymbol{p}^{k+\frac{1}{2}} - \frac{\tau}{2} \frac{\partial H}{\partial \boldsymbol{x}} \left(\boldsymbol{x}^{k+1}, \boldsymbol{p}^{k+\frac{1}{2}}\right).
\end{aligned}$$

As these equations are implicit, we must resort to fixed-point iterations. The additional information provided by the local geometry of the statistical manifold can lower the correlation between samples and increase the acceptance rate. Such an advantage will be particularly useful in high dimensions, where concentration of measure makes sampling very challenging [even though the computational cost of RMHMC will also increase as $O(d^3)$].

Recent advances have included replacing the underlying Lebesgue measure by the Hausdorff measure; as a result, H becomes $H(\boldsymbol{x}, \boldsymbol{p}) = U_{\mathcal{H}}(\boldsymbol{x}) + \frac{1}{2} \boldsymbol{p}^T \boldsymbol{G}^{-1}(\boldsymbol{x}) \boldsymbol{p}$, where $U_{\mathcal{H}}(\boldsymbol{x}) := -\log \pi_{\mathcal{H}}(\boldsymbol{x})$ is the potential energy of the target density $\pi_{\mathcal{H}}$ with respect to the Hausdorff measure. This has the advantage that the method of Hamiltonian splitting (defined in Section 3.1) can then be used to construct a geodesic integrator (Byrne & Girolami 2013). The use of Lagrangian dynamics (Fang et al. 2014, Lan et al. 2015) has also been proposed to obtain integrators that are not volume-preserving but have lower computational costs. Finally, other Riemannian metrics that do not rely on a Bayesian setting have been studied, often with the aim of improving the sampling of multimodal densities (Lan et al. 2014, Nishimura & Dunson 2016).

5. SHADOW HAMILTONIANS

We now discuss a remarkable property of symplectic integrators: the existence of a shadow Hamiltonian that is exactly conserved by the symplectic integrator (in the sense of an asymptotic expansion). The study of this quantity was inspired by backward-error analysis of differential equations and is used in molecular dynamics but is mostly unknown in the statistics literature. This is partly

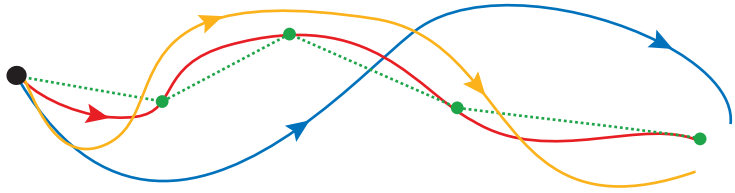

Figure 7

Shadow Hamiltonian Monte Carlo—the blue line gives the exact trajectory. The dotted green line is the numerical method and exactly follows the shadow trajectory (*red*). The orange line is the approximate shadow trajectory.

due to the geometric notions required to define it, in particular the Poisson bracket and its Lie algebra structure. In this section, we provide an intuitive introduction to the shadow Hamiltonian. This is complemented by a section in the **Supplemental Appendix**, in which we define those advanced notions more carefully.

Supplemental Material

We have seen that the leapfrog integrator and other symplectic integrators do not exactly preserve the Hamiltonian H. Over noninfinitesimal times this causes the simulated trajectory to diverge from the exact Hamiltonian trajectory. We might expect the energy along the approximate trajectory to diverge linearly with the trajectory length (number of steps). However, in practice, the energy does not diverge but merely oscillates around the correct energy even for very long trajectory lengths. The reason for this is that there is a nearby Hamiltonian that is exactly conserved by the discrete integrator. That is, we can find a shadow Hamiltonian $\tilde{H}_\tau$ that is constant along the simulated trajectory (see **Figure 7**). The shadow Hamiltonian is defined as an asymptotic expansion in the step-size that is exponentially accurate (for a small enough step-size). The aim of SHMC (Izaguirre & Hampton 2004) is to sample from a distribution with density close to $e^{-\tilde{H}_\tau}$ in order to improve the acceptance rate, and then correct for the fact that we are not sampling from the desired target density by reweighting.

Using the Baker–Campbell–Hausdorff formula, it is possible to build Hamiltonians that are arbitrarily close to the shadow Hamiltonian (Skeel & Hardy 2001), i.e., they satisfy $\tilde{H}_{[2d]} = \tilde{H}_\tau + O(\tau^{2d})$ (the square bracket notation indicates the order of the approximation). A difficulty is that the shadow Hamiltonian is not a sum of a kinetic and a potential term, and therefore the momentum refreshment step no longer just involves sampling from a Gaussian distribution. The SHMC algorithm instead samples from the new target density $\rho_M(\boldsymbol{x}, \boldsymbol{p}) := (1/Z_\rho)e^{-H_M(\boldsymbol{x},\boldsymbol{p})}$, defined by the Hamiltonian $H_M(\boldsymbol{x}, \boldsymbol{p}) := \max\{H(\boldsymbol{x}, \boldsymbol{p}), \tilde{H}_{[2d]}(\boldsymbol{x}, \boldsymbol{p}) - a\}$ where a is a constant parameter that bounds the allowed difference between $\tilde{H}_\tau$ and $\tilde{H}(\boldsymbol{x}, \boldsymbol{p})$ and must be tuned. The purpose of introducing this maximum is that it is bounded below by H. We can therefore generate Gaussian samples from H and then use rejection sampling [also called von Neumann's rejection method (Robert & Casella 2004)] to convert these into samples from ρ_M. When a is large and positive, H_M is essentially the same as H and we will achieve a high acceptance rate for the rejection sampler, while when a is large and negative we will approximate well the shadow but will have a low acceptance rate. Hence the tuning of a is critical. Given a current state $(\boldsymbol{x}, \boldsymbol{p}) \in T^*\mathcal{X}$, the algorithm proceeds as follows:

1. Draw a new momentum $\boldsymbol{p}'$ from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and accept it with probability

$$\frac{e^{-H_M(\boldsymbol{x},\boldsymbol{p}')}}{e^{-H(\boldsymbol{x},\boldsymbol{p}')}} = \min\left[1, \frac{\exp\left(a - \tilde{H}_{[2d]}(\boldsymbol{x}, \boldsymbol{p}')\right)}{\exp\left(-H(\boldsymbol{x}, \boldsymbol{p}')\right)}\right].$$

Repeat until a $\boldsymbol{p}'$ is accepted. This is simply rejection sampling.

2. Simulate Hamiltonian mechanics with initial phase $(\boldsymbol{x}, \boldsymbol{p}') \in T^*\mathcal{M}$ and Hamiltonian H using a symplectic time-reversible integrator. This yields a proposed configuration $(\boldsymbol{x}^*, \boldsymbol{p}^*)$, which we accept with probability $\min\left\{1, \frac{\rho(\boldsymbol{x}^*, \boldsymbol{p}^*)}{\rho(\boldsymbol{x}, \boldsymbol{p}')}\right\}$, else keep the old phase $(\boldsymbol{x}, \boldsymbol{p})$.

As before, step 1 is a momentum heat bath (Gibbs sampler) and step 2 is a molecular dynamics–MC step.

To calculate the sample average, reweighting is necessary to compensate for the fact that we are sampling from the wrong distribution. To do this, we reweight the generated samples by a factor $c_k := \exp(\tilde{H}(\boldsymbol{x}^l, \boldsymbol{p}^l) - H(\boldsymbol{x}^l, \boldsymbol{p}^l))$. The main advantage of this method is that the Metropolis acceptance rate will be much closer to one. However, the momentum refreshment step can become expensive and the variance of the sample average will be large if the factors c_k are not close to one.

There are, however, several issues surrounding SHMC. While the acceptance rate is greatly improved in the Metropolis step, SHMC samples from distributions with nonseparable Hamiltonians, so momentum sampling is more expensive. Moreover, it introduces a new parameter a to balance the acceptance rates of the two steps. Sweet et al. (2009) built a variant in which a canonical transformation (symplectomorphism) is used to change coordinates in order to get a separable Hamiltonian. Alternatively, a generalized SHMC algorithm has also been proposed (Akhmatskaya & Reich 2008).

The shadow Hamiltonian can be used to tune the parameters of HMC (Kennedy et al. 2012). The variance $\mathrm{Var}(H - \tilde{H})$ may be expressed as a function of Poisson brackets and integrator parameters, and it turns out that for extensive systems the Poisson brackets are almost constant. It follows that we can tune the parameters of complicated symmetric symplectic integrators and minimize this variance by simply measuring the appropriate Poisson brackets.

6. RECENT RESEARCH DIRECTIONS

6.1. Stochastic Gradient Markov Chain Monte Carlo

One of the major issues in the use of MCMC methods in a Bayesian context is the size of datasets. Imagine that we have i.i.d. observations $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ and are interested in the posterior density over some parameter $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ (here, we assume we have predefined some prior π_0): $\pi(\boldsymbol{x}|\boldsymbol{y}) \propto \pi_0(\boldsymbol{x}) \prod_{j=1}^n \mathcal{L}(\boldsymbol{y}_j|\boldsymbol{x})$. Clearly, if n is very large, then the posterior $\pi(\boldsymbol{x}|\boldsymbol{y})$ and the score functions $\partial_i \log \pi(\boldsymbol{x}|\boldsymbol{y})$ will be computationally expensive to evaluate, rendering MCMC costly. To tackle this issue, Welling & Teh (2011) suggested making use of small subsets of the entire dataset (called mini-batches) to compute the score functions, making this inference tractable once again. Although this methodology was originally developed for MALA, it was later extended to HMC algorithms (Chen et al. 2014, 2015; Ma et al. 2015). It is, however, important to note that these algorithms are not exact (in the sense that they only target an approximate target density), and the bias could be large and very difficult to assess a priori (Teh et al. 2014, Vollmer et al. 2015, Betancourt 2015).

6.2. Infinite-Dimensional Hamiltonian Monte Carlo

Recently, Beskos et al. (2011) and Cotter et al. (2013) proposed to deal with the deteriorating performance of HMC in very high dimensions by building a HMC algorithm that samples from a measure μ on an infinite-dimensional Hilbert space, such that our target measure π is a finite-dimensional projection of μ. Informally, we can think of infinite-dimensional probability distributions as being a distribution on functions (e.g., Gaussian processes or Dirichlet processes).

Measures of this form appear in a wide range of applications, from fluid dynamics to computational tomography; and more generally in Bayesian inverse problems (Stuart 2010, Beskos et al. 2016).

The algorithm samples from a measure μ on a separable Hilbert space $\mathcal{H}$, which is defined by its Radon–Nikodym derivative with respect to a dominating Gaussian measure μ_0, as given by $\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(x) \propto \exp(-\Phi(x))$ for some potential function $\Phi : \mathcal{H} \to \mathbb{R}$. Looking at HMC this way removes the dependence on the dimension d of the projection, as the algorithm is defined directly over an infinite-dimensional space. Specifically, it allows for efficient sampling from target measures in very large dimensions and the acceptance rate does not tend to 0 as $d \to \infty$, since the algorithm is well-defined in that limit.

7. CONCLUSIONS

The use of differential geometry in statistical science dates back to Rao (1945), who sought to assess the natural distance between population distributions. The Fisher–Rao metric tensor defined the Riemannian manifold structure of probability measures, and from this local manifold, geodesic distances between measures could be properly defined. This early work was then taken up by many authors, with an emphasis on studying the efficiency of statistical estimators (Efron 1982, Barndorff-Nielsen et al. 1986, Amari 1987, Critchley et al. 1993, Murray & Rice 1993). Geometry has since developed substantially and has had major impact in areas of applied statistics such as machine learning and statistical signal processing (Amari 1987).

This review has provided an accessible introduction to the necessary differential geometry, with a focus on the elements required to formally describe HMC. This should also be of interest to readers interested in the development of new methods that seek to address the growing list of challenges modern day statistical science is being called upon to address. More generally, we believe the use of geometry is essential to gain insights into more advanced methods, including shadow Hamiltonian and Riemann manifold Hamiltonian methods, and to tackle sampling issues related to the curse of dimensionality and concentration of measure in, for example, deep learning.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

A.B. was supported by a Roth scholarship from the Department of Mathematics at Imperial College London. F.X.B. was supported by the EPSRC grant [EP/L016710/1]. M.G. was supported by the EPSRC grants [EP/J016934/3, EP/K034154/1, EP/P020720/1], an EPSRC Established Career Fellowship, the EU grant [EU/259348], a Royal Society Wolfson Research Merit Award, and the Lloyds Register Foundation Programme on Data-Centric Engineering. A.D.K. was supported by STFC Consolidated Grant [ST/J000329/1]. This work was supported by The Alan Turing Institute under the EPSRC grant [EP/N510129/1]. This material was also based upon work partially supported by the National Science Foundation under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. The authors acknowledge support of the UK Defence Science and Technology Laboratory (Dstl) and Engineering and Physical Research Council (EPSRC) under grant [EP/R018413/1]. This is part of the collaboration between the US DOD, UK MOD, and UK EPSRC under the Multidisciplinary University Research Initiative.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

LITERATURE CITED

Abdulle A, Vilmart G, Zygalakis KC. 2015. Long time accuracy of Lie-Trotter splitting methods for Langevin dynamics. *SIAM J. Numer. Anal.* 53:1–16

Ajay A, Walters W, Murcko M. 1998. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Medic. Chem.* 41:3314–24

Akhmatskaya E, Reich S. 2008. GSHMC: an efficient method for molecular simulation. *J. Comp. Phys.* 227:4934–54

Amari SI. 1987. *Differential Geometrical Methods in Statistics*. New York: Springer

Arnold V. 1989. *Mathematical Methods of Classical Mechanics*. New York: Springer

Barndorff-Nielsen OE, Cox DR, Reid N. 1986. The role of differential geometry in statistical theory. *Int. Stat. Rev.* 54:83–96

Berne BJ, Straub JE. 1997. Novel methods of sampling phase space in the simulation of biological systems. *Curr. Opin. Struct. Biol.* 7:181–89

Beskos A, Girolami M, Lan S, Farrell PE, Stuart AM. 2016. Geometric MCMC for infinite-dimensional inverse problems. arXiv:1606.06351 [stat.CO]

Beskos A, Pinski FJ, Sanz-Serna JM, Stuart AM. 2011. Hybrid Monte Carlo on Hilbert spaces. *Stoch. Proc. Appl.* 121:2201–30

Beskos A, Roberts GO, Sanz-Serna JM, Stuart AM. 2013. Optimal tuning of the hybrid Monte-Carlo algorithm. *Bernoulli* 19:1501–34

Betancourt MJ. 2015. The fundamental incompatibility of Hamiltonian Monte Carlo and data subsampling. In *Proceedings of the 37th International Conference on Machine Learning (ICML 37)*, ed. F Bach, D Blei, pp. 533–40. Brookline, MA: Microtome

Betancourt MJ. 2017. A conceptual introduction to Hamiltonian Monte Carlo. arXiv:1701.02434 [stat.ME]

Betancourt MJ, Byrne S, Livingstone S, Girolami M. 2016. The geometric foundations of Hamiltonian Monte Carlo. *Bernoulli* 23:2257–98

Blanes S, Casas F, Sanz-Serna JM. 2014. Numerical integrators for the hybrid Monte Carlo method. *SIAM J. Sci. Comput.* 36:1752–69

Bui-Thanh T, Girolami M. 2014. Solving large-scale PDE-constrained Bayesian inverse problems with Riemann manifold Hamiltonian Monte Carlo. *Inverse Probl.* 30:114014

Byrne S, Girolami M. 2013. Geodesic Monte Carlo on embedded manifolds. *Scand. J. Stat.* 40:825–45

Campos CM, Sanz-Serna JM. 2015. Extra chance generalized hybrid Monte Carlo. *J. Comput. Phys.* 281:365–74

Campostrini M, Rossi P. 1990. A comparison of numerical algorithms for dynamical fermions. *Nucl. Phys. B* 329:753–64

Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, et al. 2016. Stan: a probabilistic programming language. *J. Stat. Softw.* 76:1–32

Chen C, Ding N, Carin L. 2015. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS 15)*, ed. C Cortes, DD Lee, M Sugiyama, R Garnett, pp. 2278–86. Cambridge, MA: MIT Press

Chen T, Fox EB, Guestrin C. 2014. Stochastic gradient Hamiltonian Monte Carlo. In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, ed. EP Xing, T Jebara, pp. 3663–76. Brookline, MA: Microtome

Cheung SH, Beck JL. 2009. Simulation with application to structural dynamic models with many uncertain parameters. *J. Eng. Mech.* 135:243–55

Chiu SN, Stoyan D, Kendall W, Mecke J. 2013. *Stochastic Geometry and Its Applications*. New York: Wiley

Choo K, Fleet DJ. 2001. People tracking using hybrid Monte Carlo filtering. *Proc. 8th IEEE Int. Conf. Computer Vis.*, pp. 321–28. New York: IEEE

Cotter SL, Roberts GO, Stuart A, White D. 2013. MCMC methods for functions: modifying old algorithms to make them faster. *Stat. Sci.* 28:424–46

Critchley F, Marriott P, Salmon M. 1993. Preferred point geometry and statistical manifolds. *Ann. Stat.* 21:1197–224

Diaconis P, Holmes S, Neal RM. 2000. Analysis of a nonreversible Markov chain sampler. *Ann. Appl. Prob.* 10:726–52

Dryden IL, Kent JT. 2015. *Geometry Driven Statistics*. New York: Wiley

Dryden IL, Mardia KV. 1998. *Statistical Shape Analysis*. New York: Wiley

Duane S, Kennedy AD, Pendleton BJ, Roweth D. 1987. Hybrid Monte Carlo. *Phys. Lett. B* 195:216–22

Efron B. 1982. Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Stat.* 10:1100–20

Fang Y, Sanz-Serna JM, Skeel RD. 2014. Compressible generalized hybrid Monte Carlo. *J. Chem. Phys.* 140:174108

Fernández-Pendás M, Escribano B, Radivojević T, Akhmatskaya E. 2014. Constant pressure hybrid Monte Carlo simulations in GROMACS. *J. Mol. Model.* 20:2487

Frankel T. 2012. *The Geometry of Physics*. Cambridge, UK: Cambridge Univ. Press. 3rd ed.

Fredrickson GH, Ganesan V, Drolet F. 2002. Field-theoretical computer simulation methods for polymer and complex fluids. *Macromolecules* 35:16–39

Gelfand AE, Smith AFM. 1990. Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.* 85:398–409

Girolami M, Calderhead B. 2011. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. B* 73:123–214

Hairer E, Lubich C, Wanner G. 2006. *Geometric Numerical Integration Algorithms for Ordinary Differential Equations*. New York: Springer

Hansmann UH, Okamoto Y. 1999. New Monte Carlo algorithms for protein folding. *Curr. Opin. Struct. Biol.* 9:177–83

Hastings WK. 1970. Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika* 57:97–109

Hoffman M, Gelman A. 2014. The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* 15:1593–623

Horowitz AM. 1991. A generalized guided Monte Carlo algorithm. *Phys. Lett. B* 268:247–52

Izaguirre JA, Hampton SS. 2004. Shadow hybrid Monte Carlo: an efficient propagator in phase space of macromolecules. *J. Comput. Phys.* 200:581–604

Kass RE, Vos PW. 1997. *Geometrical Foundations of Asymptotic Inference*. New York: Wiley

Kennedy AD, Silva PJ, Clark MA. 2012. Shadow Hamiltonians, Poisson brackets, and gauge theories. *Phys. Rev. D* 87:034511

Konukoglu E, Relan J, Cilingir U, Menze BH, Chinchapatnam P, et al. 2011. Efficient probabilistic model personalization integrating uncertainty on data and parameters: application to eikonal-diffusion models in cardiac electrophysiology. *Prog. Biophys. Mol. Biol.* 107:134–46

Kramer A, Calderhead B, Radde N. 2014. Hamiltonian Monte Carlo methods for efficient parameter estimation in steady state dynamical systems. *BMC Bioinform.* 15:253

Lan S, Bui-Thanh T, Christie M, Girolami M. 2016. Emulation of higher-order tensors in manifold Monte Carlo methods for Bayesian inverse problems. *J. Comput. Phys.* 308:81–101

Lan S, Stathopoulos V, Shahbaba B, Girolami M. 2015. Markov chain Monte Carlo from Lagrangian dynamics. *J. Comput. Graphic. Stat.* 24:357–78

Lan S, Streets J, Shahbaba B. 2014. Wormhole Hamiltonian Monte Carlo. *Proc. 28th AAAI Conf. Artif. Intell.*, pp. 1953–59. Palo Alto, CA: AAAI

Landau DP, Binder K. 2009. *A Guide to Monte-Carlo Simulations in Statistical Physics*. Cambridge, UK: Cambridge Univ. Press

Ledoux M. 2001. *The Concentration of Measure Phenomenon*. Providence, RI: Am. Math. Soc.

Leimkuhler B, Reich S. 2004. *Simulating Hamiltonian Dynamics*. Cambridge, UK: Cambridge Univ. Press

Livingstone S, Betancourt M, Byrne S, Girolami M. 2015. On the geometric ergodicity of Hamiltonian Monte Carlo. arXiv:1601.08057 [stat.CO]

Livingstone S, Girolami M. 2014. Information-geometric Markov chain Monte Carlo methods using diffusions. *Entropy* 16:3074–102

Ma Y, Chen T, Fox EB. 2015. A complete recipe for stochastic gradient MCMC. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS 15)*, ed. C Cortes, ND Lawrence, DD Lee, M Sugiyama, R Garnett, pp. 2917–25. Cambridge, MA: MIT Press

MacKay DJC. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK: Cambridge Univ. Press

Mehlig B, Heermann D, Forrest B. 1992. Hybrid Monte Carlo method for condensed-matter systems. *Phys. Rev. B* 45:679–85

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21:1087–92

Meyn S, Tweedie R. 1993. *Markov Chains and Stochastic Stability*. Berlin: Springer-Verlag

Murray MK, Rice JW. 1993. *Differential Geometry and Statistics*. New York: Springer

Neal RM. 2011. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, ed. S Brooks, A Gelman, GL Jones, XL Meng, pp. 113–62. Boca Raton, FL: Chapman and Hall/CRC

Nishimura A, Dunson D. 2016. Geometrically tempered Hamiltonian Monte Carlo. arXiv:1604.00872 [stat.CO]

Rao CR. 1945. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* 37:81–91

Robert C, Casella G. 2004. *Monte Carlo Statistical Methods*. New York: Springer

Robert C, Casella G. 2011. A short history of Markov chain Monte Carlo: subjective recollections from incomplete data. *Stat. Sci.* 26:102–15

Roberts GO, Rosenthal JS. 1998. Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. B* 60:255–68

Rossky PJ, Doll JD, Friedman HL. 1978. Brownian dynamics as smart Monte Carlo simulation. *J. Chem. Phys.* 69:4628–33

Scalettar R, Scalapino DJ, Sugar RL. 1986. New algorithm for the numerical simulation of fermions. *Phys. Rev. B* 34:7911–17

Schroeder KB, McElreath R, Nettle D. 2013. Variants at serotonin transporter and 2A receptor genes predict cooperative behavior differentially according to presence of punishment. *PNAS* 110:3955–60

Sen MK, Biswas R. 2017. Transdimensional seismic inversion using the reversible jump Hamiltonian Monte Carlo algorithm. *Geophysics* 82:119–34

Skeel RD, Hardy DJ. 2001. Practical construction of modified Hamiltonians. *SIAM J. Sci. Comput.* 23:1172–88

Stuart AM. 2010. Inverse problems: a Bayesian perspective. *Acta Numer.* 19:451–559

Sweet CR, Hampton SS, Skeel RD, Izaguirre JA. 2009. A separable shadow Hamiltonian hybrid Monte Carlo method. *J. Chem. Phys.* 131:174106

Teh YW, Thiery AH, Vollmer S. 2014. Consistency and fluctuations for stochastic gradient Langevin dynamics. arXiv:1409.0578 [stat.ML]

Vollmer SJ, Zygalakis KC, Teh YW. 2015. (Non-) asymptotic properties of stochastic gradient Langevin dynamics. arXiv:1501.00438 [stat.ME]

Wang Z, Mohamed S, de Freitas N. 2013. Adaptive Hamiltonian and Riemann manifold Monte Carlo samplers. In *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, ed. S Dasgupta, D McAllester, pp. 1462–70. Brookline, MA: Microtome

Welling M, Teh YW. 2011. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning (ICML 2011)*, pp. 681–88. Madison, WI: Omnipress

Yoshida H. 1990. Construction of higher order symplectic integrators. *Phys. Lett. A* 150:262–68