Entropy **2014**, *16*, 3074-3102; doi:10.3390/e16063074

OPEN ACCESS
entropy
ISSN 1099-4300
www.mdpi.com/journal/entropy

Article

Information-Geometric Markov Chain Monte Carlo Methods Using Diffusions

Samuel Livingstone [1,*] and Mark Girolami [2]

[1] Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK
[2] Department of Statistics, University of Warwick, Coventry CV4 7AL, UK;
E-Mail:m.girolami@warwick.ac.uk

* Author to whom correspondence should be addressed; E-Mail: samuel.livingstone@ucl.ac.uk;
Tel.: +44-20-7679-1872.

Received: 29 March 2014; in revised form: 23 May 2014 / Accepted: 28 May 2014 /
Published: 3 June 2014

Abstract: Recent work incorporating geometric ideas in Markov chain Monte Carlo is reviewed in order to highlight these advances and their possible application in a range of domains beyond statistics. A full exposition of Markov chains and their use in Monte Carlo simulation for statistical inference and molecular dynamics is provided, with particular emphasis on methods based on Langevin diffusions. After this, geometric concepts in Markov chain Monte Carlo are introduced. A full derivation of the Langevin diffusion on a Riemannian manifold is given, together with a discussion of the appropriate Riemannian metric choice for different problems. A survey of applications is provided, and some open questions are discussed.

Keywords: information geometry; Markov chain Monte Carlo; Bayesian inference; computational statistics; machine learning; statistical mechanics; diffusions

---

1. Introduction

There are three objectives to this article. The first is to introduce geometric concepts that have recently been employed in Monte Carlo methods based on Markov chains [1] to a wider audience. The second is to clarify what a "diffusion on a manifold" is, and how this relates to a diffusion defined on Euclidean space. Finally, we review the state-of-the-art in the field and suggest avenues for further research.

The connections between some Monte Carlo methods commonly used in statistics, physics and application domains, such as econometrics, and ideas from both Riemannian and information

geometry [2,3] were highlighted by Girolami and Calderhead [1] and the potential benefits demonstrated empirically. Two Markov chain Monte Carlo methods were introduced, the manifold Metropolis-adjusted Langevin algorithm and Riemannian manifold Hamiltonian Monte Carlo. Here, we focus on the former for two reasons. First, the intuition for why geometric ideas can improve standard algorithms is the same in both cases. Second, the foundations of the methods are quite different, and since the focus of the article is on using geometric ideas to improve performance, we considered a detailed description of both to be unnecessary. It should be noted, however, that impressive empirical evidence exists for using Hamiltonian methods in some scenarios (e.g., [4]). We refer interested readers to [5,6].

We take an expository approach, providing a review of some necessary preliminaries from Markov chain Monte Carlo, diffusion processes and Riemannian geometry. We assume only a minimal familiarity with measure-theoretic probability. More informed readers may prefer to skip these sections. We then provide a full derivation of the Langevin diffusion on a Riemannian manifold and offer some intuition for how to think about such a process. We conclude Section 4 by presenting the Metropolis-adjusted Langevin algorithm on a Riemannian manifold.

A key challenge in the geometric approach is which manifold to choose. We discuss this in Section 4.4 and review some candidates that have been suggested in the literature, along with the reasoning for each. Rather than provide a simulation study here, we instead reference studies where the methods we describe have been applied in Section 5. In Section 6, we discuss several open questions, which we feel could be interesting areas of further research and of interest to both theorists and practitioners.

Throughout, $\pi(\cdot)$ will refer to an n-dimensional probability distribution and $\pi(x)$ its density with respect to the Lebesgue measure.

2. Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a set of methods for drawing samples from a distribution, $\pi(\cdot)$, defined on a measurable space $(\mathcal{X}, \mathcal{B})$, whose density is only known up to some proportionality constant. Although the i-th sample is dependent on the $(i-1)$-th, the Ergodic Theorem ensures that for an appropriately constructed Markov chain with invariant distribution $\pi(\cdot)$, long-run averages are consistent estimators for expectations under $\pi(\cdot)$. As a result, MCMC methods have proven useful in Bayesian statistical inference, where often, the posterior density $\pi(x|y) \propto f(y|x)\pi_0(x)$ for some parameter, x (where $f(y|x)$ denotes the likelihood for data y and $\pi_0(x)$ the prior density), is only known up to a constant [7]. Here, we briefly introduce some concepts from general state space Markov chain theory together with a short overview of MCMC methods. The exposition follows [8].

2.1. Markov Chain Preliminaries

A time-homogeneous Markov chain, $\{X_m\}_{m\in\mathbb{N}}$, is a collection of random variables, X_m, each of which is defined on a measurable space $(\mathcal{X}, \mathcal{B})$, such that:

$$\mathbb{P}[X_m \in A | X_0 = x_0, ..., X_{m-1} = x_{m-1}] = \mathbb{P}[X_m \in A | X_{m-1} = x_{m-1}], \tag{1}$$

for any $A \in \mathcal{B}$. We define the transition kernel $P(x_{m-1}, A) = \mathbb{P}[X_m \in A | X_{m-1} = x_{m-1}]$ for the chain to be a map for which $P(x, \cdot)$ defines a distribution over $(\mathcal{X}, \mathcal{B})$ for any $x \in \mathcal{X}$, and $P(\cdot, A)$ is measurable

for any $A \in \mathcal{B}$. Intuitively, P defines a map from points to distributions in $\mathcal{X}$. Similarly, we define the m-step transition kernel to be:

$$P^m(x_0, A) = \mathbb{P}[X_m \in A | X_0 = x_0]. \tag{2}$$

We call a distribution $\pi(\cdot)$ invariant for $\{X_m\}_{m\in\mathbb{N}}$ if:

$$\pi(A) = \int_{\mathcal{X}} P(x, A)\pi(dx) \tag{3}$$

for all $A \in \mathcal{B}$. If $P(x, \cdot)$ admits a density, $p(x'|x)$, this can be equivalently written:

$$\pi(x') = \int_{\mathcal{X}} \pi(x)p(x'|x)dx. \tag{4}$$

The connotation of Equations (3) and (4) is that if $X_m \sim \pi(\cdot)$, then $X_{m+s} \sim \pi(\cdot)$ for any $s \in \mathbb{N}$. In this instance, we say the chain is "at stationarity". Of interest to us will be Markov chains for which there is a unique invariant distribution, which is also the limiting distribution for the chain, meaning that for any $x_0 \in \mathcal{X}$ for which $\pi(x_0) > 0$:

$$\lim_{m\to\infty} P^m(x_0, A) = \pi(A) \tag{5}$$

for any $A \in \mathcal{B}$. Certain conditions are required for Equation (5) to hold, but for all Markov chains presented here, these are satisfied (though, see [8]).

A useful condition, which is sufficient (though not necessary) for $\pi(\cdot)$ to be an invariant distribution, is reversibility, which can be shown by the relation:

$$\pi(x)p(x'|x) = \pi(x')p(x|x'). \tag{6}$$

Integrating over both sides with respect to x, we recover Equation (4). In other words, a chain is reversible if, at stationarity, the probability that $x_i \in A$ and $x_{i+1} \in B$ are equal to the probability that $x_{i+1} \in A$ and $x_i \in B$. The relation (6) will be the primary tool used to construct Markov chains with a desired invariant distribution in the next section.

2.1.1. Monte Carlo Estimates from Markov Chains

Of most interest here are estimators constructed from a Markov chain. The Ergodic Theorem states that for any chain, $\{X_m\}_{m\in\mathbb{N}}$, satisfying Equation (5) and any $g \in L^1(\pi)$, we have that:

$$\lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} g(X_i) = \mathbb{E}_\pi[g(X)] \tag{7}$$

with probability one [7]. This is a Markov chain analogue to the Law of large numbers.

The efficiency of estimators of the form $\hat{t}_m = \sum_i g(X_i)/m$ can be assessed through the autocorrelation between elements in the chain. We will assess the efficiency of $\hat{t}_m$ relative to estimators $\bar{t}_m = \sum_i g(Z_i)/m$, where $\{Z_i\}_{m\in\mathbb{N}}$ is a sequence of independent random variables, each having distribution $\pi(\cdot)$. Provided $\mathrm{Var}_\pi[g(Z_i)] < \infty$, then $\mathrm{Var}[\bar{t}_m] = \mathrm{Var}_\pi[g(Z_i)]/m$. We now seek a similar result for estimators of the form, $\hat{t}_m$.

It follows directly from the Kipnis–Varadhan Theorem [9] that an estimator, $\hat{t}_m$, from a reversible Markov chain for which $X_0 \sim \pi(\cdot)$ satisfies:

$$\lim_{m\to\infty} \frac{\text{Var}[\hat{t}_m]}{\text{Var}[\bar{t}_m]} = 1 + 2\sum_{i=1}^{\infty} \rho^{(0,i)} = \tau, \tag{8}$$

provided that $\sum_{i=1}^{\infty} i|\rho^{(0,i)}| < \infty$, where $\rho^{(0,i)} = \text{Corr}_\pi[g(X_0), g(X_i)]$. We will refer to the constant, τ, as the autocorrelation time for the chain.

Equation (8) implies that for large enough m, $\text{Var}[\hat{t}_m] \approx \tau \text{Var}[\bar{t}_m]$. In practical applications, the sum in Equation (8) is truncated to the first $p-1$ realisations of the chain, where p is the first instance at which $|\rho^{(0,p)}| < \epsilon$ for some $\epsilon > 0$. For example, in the Convergence Diagnosis and Output Analysis for MCMC (CODA) package within the R statistical software $\epsilon = 0.05$ [10,11].

Another commonly used measure of efficiency is the effective sample size $m_{eff} = m/\tau$, which gives the number of independent samples from $\pi(\cdot)$ needed to give an equally efficient estimate for $\mathbb{E}_\pi[g(X)]$. Clearly, minimising τ is equivalent to maximising m_{eff}.

The measures arising from Equation (8) give some intuition for what sort of Markov chain gives rise to efficient estimators. However, in practice, the chain will never be at stationarity. Therefore, we also assess Markov chains according to how far away they are from this point. For this, we need to measure how close $P^m(x_0, \cdot)$ is from $\pi(\cdot)$, which requires a notion of distance between probability distributions.

Although there are several appropriate choices [12], a common option in the Markov chain literature is the total variation distance:

$$\|\mu(\cdot) - \nu(\cdot)\|_{TV} := \sup_{A\in\mathcal{B}} |\mu(A) - \nu(A)|, \tag{9}$$

which informally gives the largest possible difference between the probabilities of a single event in $\mathcal{B}$ according to $\mu(\cdot)$ and $\nu(\cdot)$. If both distributions admit densities, Equation (9) can be written (see Appendix A):

$$\|\mu(\cdot) - \nu(\cdot)\|_{TV} = \frac{1}{2}\int_{\mathcal{X}} |\mu(x) - \nu(x)| dx. \tag{10}$$

which is proportional to the L_1 distance between $\mu(x)$ and $\nu(x)$. Our metric, $\|\cdot\|_{TV} \in [0,1]$, with $\|\cdot\|_{TV} = 1$ for distributions with disjoint supports and $\|\mu(\cdot) - \nu(\cdot)\|_{TV} = 0$, implies $\mu(\cdot) \equiv \nu(\cdot)$.

Typically, for an unbounded $\mathcal{X}$, the distance $\|P^m(x_0, \cdot) - \pi(\cdot)\|_{TV}$ will depend on x_0 for any finite m. Therefore, bounds on the distance are often sought via some inequality of the form:

$$\|P^m(x_0, \cdot) - \pi(\cdot)\|_{TV} \leq MV(x_0)f(m), \tag{11}$$

for some $M < \infty$, where $V : \mathcal{X} \to [1, \infty)$ depends on x_0 and is called a drift function, and $f : \mathbb{N} \to [0, \infty)$ depends on the number of iterations, m (and is often defined, such that $f(0) = 1$).

A Markov chain is called geometrically ergodic if $f(m) = r^m$ in Equation (11) for some $0 < r < 1$. If in addition to this, V is bounded above, the chain is called uniformly ergodic. Intuitively, if either condition holds, then the distribution of X_m will converge to $\pi(\cdot)$ geometrically quickly as m grows, and in the uniform case, this rate is independent of x_0. As well as providing some (often qualitative if M and r are unknown) bounds on the convergence rate of a Markov chain, geometric ergodicity implies that a central limit theorem exists for estimators of the form, $\hat{t}_m$. For more detail on this, see [13,14].

In practice several approximate methods also exist to assess whether a chain is close enough to stationarity for long-run averages to provide suitable estimators (e.g., [15]). The MCMC practitioner also uses a variety of visual aids to judge whether an estimate from the chain will be appropriate for his or her needs.

2.2. Markov Chain Monte Carlo

Now that we have introduced Markov chains, we turn to simulating them. The objective here is to devise a method for generating a Markov chain, which has a desired limiting distribution, $\pi(\cdot)$. In addition, we would strive for the convergence rate to be as fast as possible and the effective sample size to be suitably large relative to the number of iterations. Of course, the computational cost of performing an iteration is also an important practical consideration. Ideally, any method would also require limited problem-specific alterations, so that practitioners are able to use it with as little knowledge of the inner workings as is practical.

Although other methods exist for constructing chains with a desired limiting distribution, a popular choice is the Metropolis–Hastings algorithm [7]. At iteration i, a sample is drawn from some candidate transition kernel, $Q(x_{i-1}, \cdot)$, and then either accepted or rejected (in which case, the state of the chain remains x_{i-1}). We focus here on the case where $Q(x_{i-1}, \cdot)$ admits a density, $q(x'|x_{i-1})$, for all $x_{i-1} \in \mathcal{X}$ (though, see [8]). In this case, a single step is shown below (the wedge notation $a \wedge b$ denotes the minimum of a and b). The "acceptance rate", $\alpha(x_{i-1}, x')$, governs the behaviour of the chain, so that, when it is close to one, then many proposed moves are accepted, and the current value in the chain is constantly changing. If it is on average close to zero, then many proposals are rejected, so that the chain will remain in the same place for many iterations. However, $\alpha \approx 1$ is typically not ideal, often resulting in a large autocorrelation time (see below). The challenge in practice is to find the right acceptance rate to balance these two extremes.

Algorithm 1 Metropolis–Hastings, single iteration.

Require: x_{i-1}
  Draw $X' \sim Q(x_{i-1}, \cdot)$
  Draw $Z \sim U[0, 1]$
  Set $\alpha(x_{i-1}, x') \leftarrow 1 \wedge \frac{\pi(x')q(x_{i-1}|x')}{\pi(x_{i-1})q(x'|x_{i-1})}$
  if $z < \alpha(x_{i-1}, x')$ **then**
    Set $x_i \leftarrow x'$
  else
    Set $x_i \leftarrow x_{i-1}$
  end if

Combining the "proposal" and "acceptance" steps, the transition kernel for the resulting Markov chain is:

$$P(x, A) = r(x)\delta_x(A) + \int_A \alpha(x, x')q(x'|x)dx', \tag{12}$$

for any $A \in \mathcal{B}$, where:

$$r(x) = 1 - \int_{\mathcal{X}} \alpha(x, x')q(x'|x)dx'$$

is the average probability that a draw from $Q(x, \cdot)$ will be rejected, and $\delta_x(A) = 1$ if $x \in A$ and zero, otherwise. A Markov chain defined in this way will have $\pi(\cdot)$ as an invariant distribution, since the chain is reversible for $\pi(\cdot)$. We note here that:

$$\begin{aligned}\pi(x_{i-1})q(x_i|x_{i-1})\alpha(x_{i-1}, x_i) &= \pi(x_{i-1})q(x_i|x_{i-1}) \wedge \pi(x_i)q(x_{i-1}|x_i) \\ &= \alpha(x_i, x_{i-1})q(x_{i-1}|x_i)\pi(x_i)\end{aligned}$$

in the case that the proposed move is accepted and that if the proposed move is rejected, then $x_i = x_{i-1}$; so the chain is reversible for $\pi(\cdot)$. It can be shown that $\pi(\cdot)$ is also the limiting distribution for the chain [7].

The convergence rate and autocorrelation time of a chain produced by the algorithm are dependent on both the choice of proposal, $Q(x_{i-1}, \cdot)$, and the target distribution, $\pi(\cdot)$. For simple forms of the latter, less consideration is required when choosing the former. A broad objective among researchers in the field is to find classes of proposal kernels that produce chains that converge and mix quickly for a large class of target distributions. We first review a simple choice before discussing one that is more sophisticated, and the will be the focus of the rest of the article.

2.3. Random Walk Proposals

An extremely simple choice for $Q(x, \cdot)$ is one for which:

$$q(x'|x) = q(\|x' - x\|) \tag{13}$$

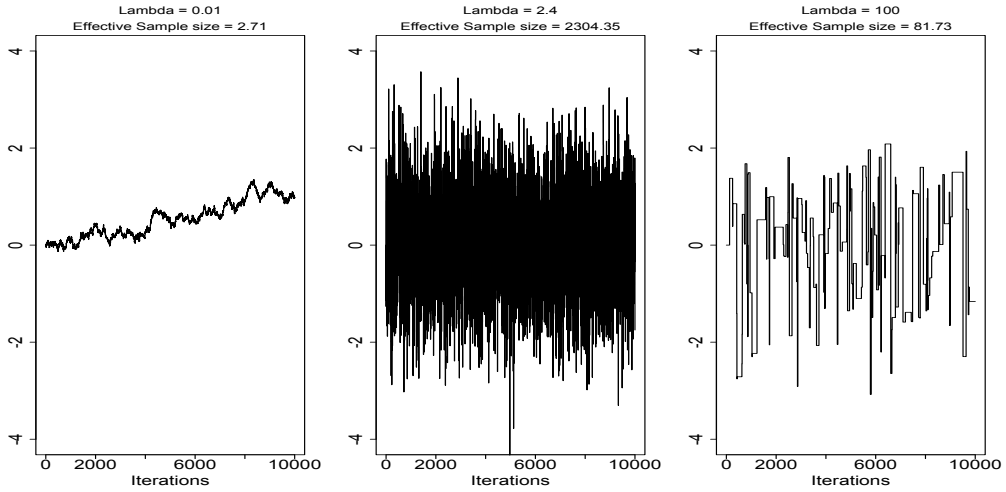
where $\|\cdot\|$ denotes some appropriate norm on $\mathcal{X}$, meaning the proposal is symmetric. In this case, the acceptance rate reduces to:

$$\alpha(x, x') = 1 \wedge \frac{\pi(x')}{\pi(x)}. \tag{14}$$

In addition to simplifying calculations, Equation (14) strengthens the intuition for the method, since proposed moves with higher density under $\pi(\cdot)$ will always be accepted. A typical choice for $Q(x, \cdot)$ is $\mathcal{N}(x, \lambda^2\Sigma)$, where the matrix, Σ, is often chosen in an attempt to match the correlation structure of $\pi(\cdot)$ or simply taken as the identity [16]. The tuning parameter, λ, is the only other user-specific input required.

Much research has been conducted into properties of the random walk Metropolis algorithm (RWM). It has been shown that the optimal acceptance rate for proposals tends to 0.234 as the dimension, n, of the state space, $\mathcal{X}$, tends to ∞ for a wide class of targets (e.g., [17,18]). The intuition for an optimal acceptance rate is to find the right balance between the distance of proposed moves and the chances of acceptance. Increasing the former will reduce the autocorrelation in the chain if the proposal is accepted, but if it is rejected, the chain will not move at all, so autocorrelation will be high. Random walk proposals are sometimes referred to as blind (e.g., [19]), as no information about $\pi(\cdot)$ is used when generating proposals, so typically, very large moves will result in a very low chance of acceptance, while small moves will be accepted, but result in very high autocorrelation for the chain. Figure 1 demonstrates this in the simple case where $\pi(\cdot)$ is a one-dimensional $\mathcal{N}(0, 1^2)$ distribution.

Figure 1. These traceplots show the evolution of three RWM Markov chains for which $\pi(\cdot)$ is a $\mathcal{N}(0, 1^2)$ distribution, with different choices for λ.



Several authors have also shown that for certain classes of $\pi(\cdot)$, the tuning parameter, λ, should be chosen, such that $\lambda^2 \propto n^{-1}$, so that $\alpha \nrightarrow 0$ as $n \to \infty$ [20]. Because of this, we say that algorithm efficiency "scales" $O(n^{-1})$ as the dimension n of $\pi(\cdot)$ increases.

Ergodicity results for a Markov chain constructed using the RWM algorithm also exist [21–23]. At least exponentially light tails are a necessity for $\pi(x)$ for geometric ergodicity, which means that $\pi(x)/e^{-\|x\|} \to c$ as $\|x\| \to \infty$, for some constant, c. For super-exponential tails (where $\pi(x) \to 0$ at a faster than the exponential rate), additional conditions are required [21,23]. We demonstrate with a simple example why heavy-tailed forms of $\pi(x)$ pose difficulties here (where $\pi(x) \to 0$ at a rate slower than $e^{-\|x\|}$).

Example: Take $\pi(x) \propto 1/(1 + x^2)$, so that $\pi(\cdot)$ is a Cauchy distribution. Then, if $X' \sim \mathcal{N}(x, \lambda^2)$, the ratio $\pi(x')/\pi(x) = (1 + x^2)/(1 + (x')^2) \to 1$ as $|x| \to \infty$. Therefore, if x_0 is far away from zero, the Markov chain will dissolve into a random walk, with almost every proposal being accepted.

It should be noted that starting the chain from at or near zero can also cause problems in the above example, as the tails of the distribution may not be explored. See [7] for more detail here.

Ergodicity results for the RWM also exist for specific classes of the statistical model. Conditions for geometric ergodicity in the case of generalised linear mixed models are given in [24], while spherically constrained target densities are discussed in [25]. In [26], the authors provide necessary conditions for the geometric convergence of RWM algorithms, which are related to the existence of exponential moments for $\pi(\cdot)$ and $P(x, \cdot)$. Weaker forms of ergodicity and corresponding conditions are also discussed in the paper.

In the remainder of the article, we will primarily discuss another approach to choosing Q, which has been shown empirically [1] and, in some cases, theoretically [20] to be superior to the RWM algorithm, though it should be noted that random walk proposals are still widely used in practice and are often sufficient for more straightforward problems [16].

3. Diffusions

In MCMC, we are concerned with discrete time processes. However, often, there are benefits to first considering a continuous time process with the properties we desire. For example, some continuous time processes can be specified via a form of differential equation. In this section, we derive a choice for a Metropolis–Hastings proposal kernel based on approximations to diffusions, those continuous-time n-dimensional Markov processes $(X_t)_{t\geq 0}$ for which any sample path $t \mapsto X_t(\omega)$ is a continuous function with probability one. For any fixed t, we assume X_t is a random variable taking values on the measurable space $(\mathcal{X}, \mathcal{B})$ as before. The motivation for this section is to define a class of diffusions for which $\pi(\cdot)$ is the invariant distribution. First, we provide some preliminaries, followed by an introduction to our main object of study, the Langevin diffusion.

3.1. Preliminaries

We focus on the class of time-homogeneous Itô diffusions, whose dynamics are governed by a stochastic differential equation of the form:

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t, \quad X_0 = x_0, \tag{15}$$

where $(B_t)_{t\geq 0}$ is a standard Brownian motion and the drift vector, b, and volatility matrix, σ, are Lipschitz continuous [27]. Since $\mathbb{E}[B_{t+\triangle t} - B_t | B_t = b_t] = 0$ for any $\triangle t \geq 0$, informally, we can see that:

$$\mathbb{E}[X_{t+\triangle t} - X_t | X_t = x_t] = b(x_t)\triangle t + o(\triangle t), \tag{16}$$

implying that the drift dictates how the mean of the process changes over a small time interval, and if we define the process $(M_t)_{t\geq 0}$ through the relation:

$$M_t = X_t - \int_0^t b(X_s)ds \tag{17}$$

then we have:

$$\mathbb{E}[(M_{t+\triangle t} - M_t)(M_{t+\triangle t} - M_t)^T | M_t = m_t, X_t = x_t] = \sigma(x_t)\sigma(x_t)^T \triangle t + o(\triangle t), \tag{18}$$

giving the stochastic part of the relationship between $X_{t+\triangle t}$ and X_t for small enough $\triangle t$; see, e.g., [28].

While Equation(15) is often a suitable description of an Itô diffusion, it can also be characterised through an infinitesimal generator, $\mathcal{A}$, which describes how functions of the process are expected to evolve. We define this partial differential operator through its action on a function, $f \in C_0(\mathcal{X})$, as:

$$\mathcal{A}f(X_t) = \lim_{\triangle t \to 0} \frac{\mathbb{E}[f(X_{t+\triangle t}) | X_t = x_t] - f(x_t)}{\triangle t}, \tag{19}$$

though $\mathcal{A}$ can be associated with the drift and volatility of $(X_t)_{t\geq 0}$ by the relation:

$$\mathcal{A}f(x) = \sum_i b_i(x)\frac{\partial f}{\partial x_i}(x) + \frac{1}{2}\sum_{i,j} V_{ij}(x)\frac{\partial^2 f}{\partial x_i \partial x_j}(x), \tag{20}$$

where $V_{ij}(x)$ denotes the component in row i and column j of $\sigma(x)\sigma(x)^T$ [27].

As in the discrete case, we can describe the transition kernel of a continuous time Markov process, $P^t(x_0, \cdot)$. In the case of an Itô diffusion, $P^t(x_0, \cdot)$ admits a density, $p_t(x|x_0)$, which, in fact, varies smoothly as a function of t. The Fokker–Planck equation describes this variation in terms of the drift and volatility and is given by:

$$\frac{\partial}{\partial t} p_t(x|x_0) = -\sum_i \frac{\partial}{\partial x_i}[b_i(x)p_t(x|x_0)] + \frac{1}{2}\sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j}[V_{ij}(x)p_t(x|x_0)]. \tag{21}$$

Although, typically, the form of $P^t(x_0, \cdot)$ is unknown, the expectation and variance of $X_t \sim P^t(x_0, \cdot)$ are given by the integral equations:

$$\mathbb{E}[X_t|X_0 = x_0] = x_0 + \mathbb{E}\left[\int_0^t b(X_s)ds\right],$$
$$\mathbb{E}[(X_t - \mathbb{E}[X_t])(X_t - \mathbb{E}[X_t])^T|X_0 = x_0] = \mathbb{E}\left[\int_0^t \sigma(X_s)\sigma(X_s)^T ds\right],$$

where the second of these is a result of the Itô isometry [27]. Continuing the analogy, a natural question is whether a diffusion process has an invariant distribution, $\pi(\cdot)$, and whether:

$$\lim_{t\to\infty} P^t(x_0, A) = \pi(A) \tag{22}$$

for any $A \in \mathcal{B}$ and any $x_0 \in \mathcal{X}$, in some sense. For a large class of diffusions (which we confine ourselves to), this is, in fact, the case. Specifically, in the case of positive Harris recurrent diffusions with invariant distribution $\pi(\cdot)$, all compact sets must be small for some skeleton chain, see [29] for details. In addition, Equation (21) provides a means of finding $\pi(\cdot)$, given b and σ. Setting the left-hand side of Equation (21) to zero gives:

$$\sum_i \frac{\partial}{\partial x_i}[b_i(x)\pi(x)] = \frac{1}{2}\sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j}[V_{ij}(x)\pi(x)], \tag{23}$$

which can be solved to find $\pi(\cdot)$.

3.2. Langevin Diffusions

Given Equation (23), our goal becomes clearer: find drift and volatility terms, so that the resulting dynamics describe a diffusion, which converges to some user-defined invariant distribution, $\pi(\cdot)$. This process can then be used as a basis for choosing Q in a Metropolis–Hastings algorithm. The Langevin diffusion, first used to describe the dynamics of molecular systems [30], is such a process, given by the solution to the stochastic differential equation:

$$dX_t = \frac{1}{2}\nabla \log \pi(X_t)dt + dB_t, \quad X_0 = x_0. \tag{24}$$

Since $V_{ij}(x) = \mathbb{1}_{\{i=j\}}$, it is clear that

$$\frac{1}{2}\frac{\partial}{\partial x_i}[\log \pi(x)]\pi(x) = \frac{1}{2}\frac{\partial}{\partial x_i}\pi(x), \quad \forall i, \tag{25}$$

which is a sufficient condition for Equation (23) to hold. Therefore, for any case in which $\pi(x)$ is suitably regular, so that $\nabla \log \pi(x)$ is well-defined and the derivatives in Equation (23) exist, we can use (24) to construct a diffusion, which has invariant distribution, $\pi(\cdot)$.

Roberts and Tweedie [31] give sufficient conditions on $\pi(\cdot)$ under which a diffusion, $(X_t)_{t\geq 0}$, with dynamics given by Equation (24), will be ergodic, meaning:

$$\|P^t(x_0, \cdot) - \pi(\cdot)\|_{TV} \to 0 \tag{26}$$

as $t \to \infty$, for any $x_0 \in \mathcal{X}$.

3.3. Metropolis-Adjusted Langevin Algorithm

We can use Langevin diffusions as a basis for MCMC in many ways, but a popular variant is known as the Metropolis-adjusted Langevin algorithm (MALA), whereby $Q(x, \cdot)$ is constructed through a Euler–Maruyama discretisation of (24) and used as a candidate kernel in a Metropolis–Hastings algorithm. The resulting Q is:

$$Q(x, \cdot) \equiv \mathcal{N}\left(x + \frac{\lambda^2}{2}\nabla \log \pi(x), \lambda^2 I\right), \tag{27}$$

where λ is again a tuning parameter.

Before we discuss the theoretical properties of the approach, we first offer an intuition for the dynamics. From Equation (27), it can be seen that Langevin-type proposals comprise a deterministic shift towards a local mode of $\pi(x)$, combined with some random additive Gaussian noise, with variance λ^2 for each component. The relative weights of the deterministic and random parts are fixed, given as they are by the parameter, λ. Typically, if $\lambda^{1/2} \gg \lambda$, then the random part of the proposal will dominate and *vice versa* in the opposite case, though this also depends on the form of $\nabla \log \pi(x)$ [31].

Again, since this is a Metropolis–Hastings method, choosing λ is a balance between proposing large enough jumps and ensuring that a reasonable proportion are accepted. It has been shown that in the limit, as $n \to \infty$, the optimal acceptance rate for the algorithm is 0.574 [20] for forms of $\pi(\cdot)$, which either have independent and identically distributed components or whose components only differ by some scaling factor [20]. In these cases, as $n \to \infty$, the parameter, λ, must be $\propto n^{-1/3}$, so we say the algorithm efficiency scales $O(n^{-1/3})$. Note that these results compare favourably with the $O(n^{-1})$ scaling of the random walk algorithm.

Convergence properties of the method have also been established. Roberts and Tweedie [31] highlight some cases in which MALA is either geometrically ergodic or not. Typically, results are based on the tail behaviour of $\pi(x)$. If these tails are heavier than exponential, then the method is typically not geometrically ergodic and similarly if the tails are lighter than Gaussian. However, in the in between case, the converse is true. We again offer two simple examples for intuition here.

Example: Take $\pi(x) \propto 1/(1 + x^2)$ as in the previous example. Then, $\nabla \log \pi(x) = -2x/(1 + x^2)^2 \to 0$ as $|x| \to \infty$. Therefore, if x_0 is far away from zero, then the MALA will be approximately equal to the RWM algorithm and, so, will also dissolve into a random walk.

Example: Take $\pi(x) \propto e^{-x^4}$. *Then,* $\nabla \log \pi(x) = -4x^3$ *and* $X' \sim \mathcal{N}(x - 4\lambda^2 x^3, \lambda^2)$. *Therefore, for any fixed* λ, *there exists* $c > 0$, *such that, for* $|x_0| > c$, *we have* $|4\lambda^2 x^3| >> x$ *and* $|x - 4\lambda^2 x^3| >> \lambda$, *suggesting that MALA proposals will quickly spiral further and further away from any neighbourhood of zero, and hence, nearly all will be rejected.*

For cases where there is a strong correlation between elements of x or each element has a different marginal variance, the MALA can also be "pre-conditioned" in a similar way to the RWM, so that the covariance structure of proposals more accurately reflects that of $\pi(x)$ [32]. In this case, proposals take the form:

$$Q(x, \cdot) \equiv \mathcal{N}\left(x + \frac{\lambda^2}{2}\Sigma \nabla \log \pi(x), \lambda^2 \Sigma\right), \tag{28}$$

where λ is again a tuning parameter. It can be shown that provided Σ is a constant matrix, $\pi(x)$ is still the invariant distribution for the diffusion on which Equation (28) is based [33].

4. Geometric Concepts in Markov Chain Monte Carlo

Ideas from information geometry have been successfully applied to statistics from as early as [34]. More widely, other geometric ideas have also been applied, offering new insight into common problems (e.g., [35,36]). A survey is given in [37]. In this section, we suggest why some ideas from differential geometry may be beneficial for sampling methods based on Markov chains. We then review what is meant by a "diffusion on a manifold", before turning to the specific case of Equation (24). After this, we discuss what can be learned from work in information geometry in this context.

4.1. Manifolds and Markov Chains

We often make assumptions in MCMC about the properties of the space, $\mathcal{X}$, in which our Markov chains evolve. Often $\mathcal{X} = \mathbb{R}^n$ or a simple re-parametrisation would make it so. However, here, $\mathbb{R}^n = \{(a_1, ..., a_n) : a_i \in (-\infty, \infty)\ \forall i\}$. The additional assumption that is often made is that $\mathbb{R}^n$ is Euclidean, an inner product space with the induced distance metric:

$$d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}. \tag{29}$$

For sampling methods based on Markov chains that explore the space locally, like the RWM and MALA, it may be advantageous to instead impose a different metric structure on the space, $\mathcal{X}$, so that some points are drawn closer together and others pushed further apart. Intuitively, one can picture distances in the space being defined, such that if the current position in the chain is far from an area of $\mathcal{X}$, which is "likely to occur" under $\pi(\cdot)$, then the distance to such a typical set could be reduced. Similarly, once this region is reached, the space could be "stretched" or "warped", so that it is explored as efficiently as possible.

While the idea is attractive, it is far from a constructive definition. We only have the pre-requisite that $(\mathcal{X}, d)$ must be a metric space. However, as Langevin dynamics use gradient information, we will require $(\mathcal{X}, d)$ to be a space on which we can do differential calculus. Riemannian manifolds are an appropriate choice, therefore, as the rules of differentiation are well understand for functions defined on

them [38,39], while we are still free to define a more local notion of distance than Euclidean. In this section, we write $\mathbb{R}^n$ to denote the Euclidean vector space.

4.2. Preliminaries

We do not provide a full overview of Riemannian geometry here [38–40]. We simply note that for our purposes, we can consider an n-dimensional Riemannian manifold (henceforth, manifold) to be an n-dimensional metric space, in which distances are defined in a specific way. We also only consider manifolds for which a global coordinate chart exists, meaning that a mapping $r : \mathbb{R}^n \to M$ exists, which is both differentiable and invertible and for which the inverse is also differentiable (a diffeomorphism). Although this restricts the class of manifolds available (the sphere, for example, is not in this class), it is again suitable for our needs and avoids the practical challenges of switching between coordinate patches. The connection with $\mathbb{R}^n$ defined through r is crucial for making sense of differentiability in M. We say a function $f : M \to \mathbb{R}$ is "differentiable" if $(f \circ r) : \mathbb{R}^n \to \mathbb{R}$ is [39].

As has been stated, Equation (29) can be induced via a Euclidean inner product, which we denote $\langle \cdot, \cdot \rangle$. However, it will aid intuition to think of distances in $\mathbb{R}^n$ via curves:

$$\gamma : [0, 1] \to \mathbb{R}^n. \tag{30}$$

We could think of the distance between two points in $x, y \in \mathbb{R}^n$ as the minimum length among all curves that pass through x and y. If $\gamma(0) = x$ and $\gamma(1) = y$, the length is defined as:

$$L(\gamma) = \int_0^1 \sqrt{\langle \gamma'(t), \gamma'(t) \rangle} dt, \tag{31}$$

giving the metric:

$$d(x, y) = \inf \{L(\gamma) : \gamma(0) = x, \gamma(1) = y\}. \tag{32}$$

In $\mathbb{R}^n$, the curve with a minimum length will be a straight line, so that Equation (32) agrees with Equation (29). More generally, we call a solution to Equation (32) a geodesic [38].

In a vector space, metric properties can always be induced through an inner product (which also gives a notion of orthogonality). Such a space can be thought of as "flat", since for any two points, y and z, the straight line $ay + (1 - a)z,\ a \in [0, 1]$ is also contained in the space. In general, manifolds do not have vector space structure globally, but do so at the infinitesimal level. As such, we can think of them as "curved". We cannot always define an inner product, but we can still define distances through (32). We define a curve on a manifold, M, as $\gamma_M : [0, 1] \to M$. At each point $\gamma_M(t) = p \in M$, the velocity vector, $\gamma'_M(t)$, lies in an n-dimensional vector space, which touches M at p. These are known as tangent spaces, denoted T_pM, which can be thought of as local linear approximations to M. We can define an inner product on each as $g_p : T_pM \to \mathbb{R}$, which allows us to define a generalisation of (31) as:

$$L(\gamma_M) = \int_0^1 \sqrt{g_p(\gamma'_M(t), \gamma'_M(t))} dt. \tag{33}$$

and provides a means to define a distance metric on the manifold as $d(x, y) = \inf \{L(\gamma_M) : \gamma_M(0) = x, \gamma_M(1) = y\}$. We emphasise the difference between this distance metric on M and g_p, which is called a Riemannian metric or metric tensor and which defines an inner product on T_pM.

Embeddings and Local Coordinates

So far, we have introduced manifolds as abstract objects. In fact, they can also be considered as objects that are embedded in some higher-dimensional Euclidean space. A simple example is any two-dimensional surface, such as the unit sphere, lying in $\mathbb{R}^3$. If a manifold is embedded in this way, then metric properties can be induced from the ambient Euclidean space.

We seek to make these ideas more concrete through an example, the graph of a function, $f(x_1, x_2)$, of two variables, x_1 and x_2. The resulting map, r, is:

$$r : \mathbb{R}^2 \to M \tag{34}$$

$$r(x_1, x_2) = (x_1, x_2, f(x_1, x_2)). \tag{35}$$

We can see that M is embedded in $\mathbb{R}^3$, but that any point can be identified using only two coordinates, x_1 and x_2. In this case, each T_pM is a plane, and therefore, a two-dimensional subspace of $\mathbb{R}^3$, so: (i) it inherits the Euclidean inner product, $\langle \cdot, \cdot \rangle$; and (ii) any vector, $v \in T_pM$, can be expressed as a linear combination of any two linearly independent basis vectors (a canonical choice is the partial derivatives $\partial r / \partial x_1 := r_1$ and r_2, evaluated at $x = r^{-1}(p) \in \mathbb{R}^2$). The resulting inner product, $g_p(v, w)$, between two vectors, $v, w \in T_pM$, can be induced from the Euclidean inner product as:

$$\begin{aligned}
\langle v, w \rangle &= \langle v_1 r_1(x) + v_2 r_2(x), w_1 r_1(x) + w_2 r_2(x) \rangle, \\
&= v_1 w_1 \langle r_1(x), r_1(x) \rangle + v_1 w_2 \langle r_1(x), r_2(x) \rangle + v_2 w_1 \langle r_2(x), r_1(x) \rangle + v_2 w_2 \langle r_2(x), r_2(x) \rangle, \\
&= v^T G(x) w,
\end{aligned}$$

where:

$$G(x) = \begin{pmatrix} \langle r_1(x), r_1(x) \rangle & \langle r_1(x), r_2(x) \rangle \\ \langle r_1(x), r_2(x) \rangle & \langle r_2(x), r_2(x) \rangle \end{pmatrix} \tag{36}$$

and we use v_i, w_i to denote the components of v and w. To write (31) using this notation, we define the curve, $x(t) \in \mathbb{R}^2$, corresponding to $\gamma_M(t) \in M$ as $x = (r^{-1} \circ \gamma_M) : [0, 1] \to \mathbb{R}^2$. Equation (31) can then be written:

$$L(\gamma_M) = \int_0^1 \sqrt{x'(t)^T G(x(t)) x'(t)} dt, \tag{37}$$

which can be used in (32) as before.

The key point is that, although we have started with an object embedded in $\mathbb{R}^3$, we can compute the Riemannian metric, $g_p(v, w)$ (and, hence, distances in M), using only the two-dimensional "local" coordinates (x_1, x_2). We also need not have explicit knowledge of the mapping, r, only the components of the positive definite matrix, $G(x)$. The Nash embedding theorem [41] in essence enables us to define manifolds by the reverse process: simply choose the matrix, $G(x)$, so that we define a metric space with suitable distance properties, and some object embedded in some higher-dimensional Euclidean space will exist for which these metric properties can be induced as above. Therefore, to define our new space, we simply choose an appropriate matrix-valued map, $G(x)$ (we discuss this choice in Section 4.4). If $G(x)$ does not depend on x, then M has a vector space structure and can be thought of as "flat". Trivially, $G(x) = I$ gives Euclidean n-space.

We can also define volumes on a Riemannian manifold in local coordinates. Following standard coordinate transformation rules, we can see that for the above example, the area element, dx, in $\mathbb{R}^2$ will change according to a Jacobian $J = |(Dr)^T(Dr)|^{1/2}$, where $Dr = \partial(p_1, p_2, p_3)/\partial(x_1, x_2)$. This reduces to $J = |G(x)|^{1/2}$, which is also the case for more general manifolds [38]. We therefore define the Riemannian volume measure on a manifold, M, in local coordinates as:

$$\text{Vol}_M(dx) = |G(x)|^{\frac{1}{2}} dx. \tag{38}$$

If $G(x) = I$, then this reduces to the Lebesgue measure.

4.3. Diffusions on Manifolds

By a "diffusion on a manifold" in local coordinates, we actually mean a diffusion defined on Euclidean space. For example, a realisation of Brownian motion on the surface, $S \subset \mathbb{R}^3$, defined in Figure 2 through $r(x_1, x_2) = (x_1, x_2, \sin(x_1) + 1)$ will be a sample path, which is defined on S and "looks locally" like Brownian motion in a neighbourhood of any point, $p \in S$. However, the pre-image of this sample path (through r^{-1}) will not be a realisation of a Brownian motion defined on $\mathbb{R}^2$, owing to the nonlinearity of the mapping. Therefore, to define "Brownian motion on S", we define some diffusion $(X_t)_{t\geq 0}$ that takes values in $\mathbb{R}^2$, for which the process $(r(X_t))_{t\geq 0}$ "looks locally" like a Brownian motion (and lies on S). See [42] for more intuition here.

Our goal, therefore, is to define a diffusion on Euclidean space, which, when mapped onto a manifold through r, becomes the Langevin diffusion described in (24) by the above procedure. Such a diffusion takes the form:

$$dX_t = \frac{1}{2}\tilde{\nabla} \log \tilde{\pi}(X_t)dt + d\tilde{B}_t, \tag{39}$$

where those objects marked with a tilde must be defined appropriately. The next few paragraphs are technical, and readers aiming to simply grasp the key points may wish to skip to the end of this Subsection.

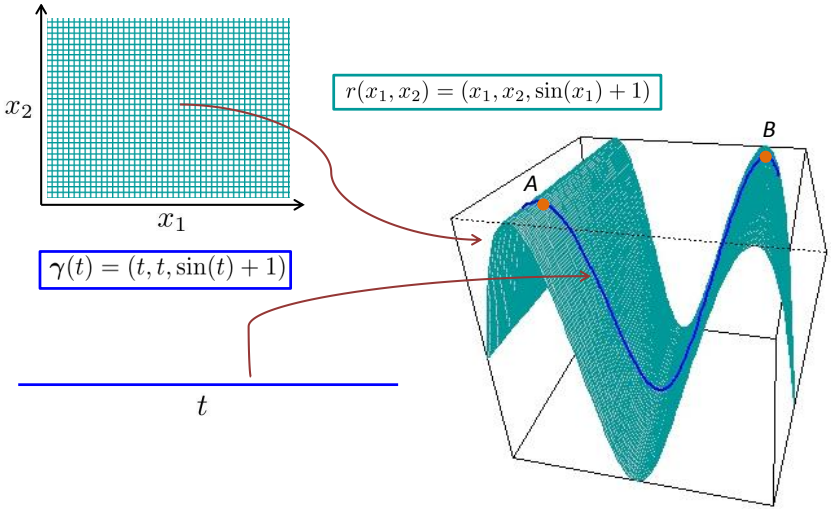
We turn first to $(\tilde{B}_t)_{t\geq 0}$, which we use to denote Brownian motion on a manifold. Intuitively, we may think of a construction based on embedded manifolds, by setting $\tilde{B}_0 = p \in M$, and for each increment sampling some random vector in the tangent space T_pM, and then moving along the manifold in the prescribed direction for an infinitesimal period of time before re-sampling another velocity vector from the next tangent space [42]. In fact, we can define such a construction using Stratonovich calculus and show that the infinitesimal generator can be written using only local coordinates [28]. Here, we instead take the approach of generalising the generator directly from Euclidean space to the local coordinates of a manifold, arriving at the same result. We then deduce the stochastic differential equation describing $(\tilde{B}_t)_{t\geq 0}$ in Itô form using (20).

For a standard Brownian motion on $\mathbb{R}^n$, $\mathcal{A} = \triangle/2$, where $\triangle$ denotes the Laplace operator:

$$\triangle f = \sum_i \frac{\partial^2 f}{\partial x_i^2} = \text{div}(\nabla f). \tag{40}$$

Substituting $\mathcal{A} = \triangle/2$ into (20) trivially gives $b_i(x) = 0\,\forall i$, $V_{ij}(x) = \mathbb{1}_{\{i=j\}}$, as required. The Laplacian, $\triangle f(x)$, is the divergence of the gradient vector field of some function, $f \in C^2(\mathbb{R}^n)$, and its value at $x \in \mathbb{R}^n$ can be thought of as the average value of f in some neighbourhood of x [43].

Figure 2. A two-dimensional manifold (surface) embedded in $\mathbb{R}^3$ through $r(x_1, x_2) = (x_1, x_2, \sin(x_1) + 1)$, parametrised by the local coordinates, x_1 and x_2. The distance between points A and B is given by the length of the curve $\gamma(t) = (t, t, \sin(t) + 1))$.



To define a Brownian motion on any manifold, the gradient and divergence must be generalised. We provide a full derivation in Appendix B, which shows that the gradient operator on a manifold can be written in local coordinates as $\nabla_M = G^{-1}(x)\nabla$. Combining with the operator, div_M, we can define a generalisation of the Laplace operator, known as the Laplace–Beltrami operator (e.g., [44,45]), as:

$$\triangle_{LB} f = \text{div}_M(\nabla_M f) = |G(x|^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \left(|G(x)|^{\frac{1}{2}} \sum_{j=1}^{n} \{G^{-1}(x)\}_{ij} \frac{\partial f}{\partial x_j} \right), \tag{41}$$

for some $f \in C_0^2(M)$.

The generator of a Brownian motion on M is $\triangle_{LB}/2$ [44]. Using (20), the resulting diffusion has dynamics given by:

$$d\tilde{B}_t = \Omega(X_t)dt + \sqrt{G^{-1}(X_t)}dB_t,$$
$$\Omega_i(X_t) = \frac{1}{2}|G(X_t)|^{-\frac{1}{2}} \sum_{j=1}^{n} \frac{\partial}{\partial x_j} \left(|G(X_t)|^{\frac{1}{2}} \{G^{-1}(X_t)\}_{ij} \right).$$

Those familiar with the Itô formula will not be surprised by the additional drift term, $\Omega(X_t)$. As Itô integrals do not follow the chain rule of ordinary calculus, non-linear mappings of martingales, such as $(B_t)_{t\geq 0}$, typically result in drift terms being added to the dynamics (e.g., [27]).

To define $\tilde{\nabla}$, we simply note that this is again the gradient operator on a general manifold, so $\tilde{\nabla} = G^{-1}(x)\nabla$. For the density, $\tilde{\pi}(x)$, we note that this density will now implicitly be defined with respect to the volume measure, $|G(x)|^{\frac{1}{2}}dx$, on the manifold. Therefore, to ensure the diffusion (39) has the correct invariant density with respect to the Lebesgue measure, we define:

$$\tilde{\pi}(x) = \pi(x)|G(x)|^{-\frac{1}{2}}. \tag{42}$$

Putting these three elements together, Equation (39) becomes:

$$dX_t = \frac{1}{2}G^{-1}(X_t)\nabla \log\left\{\pi(X_t)|G(X_t)|^{-\frac{1}{2}}\right\}dt + \Omega(X_t)dt + \sqrt{G^{-1}(X_t)}dB_t,$$

which, upon simplification, becomes:

$$dX_t = \frac{1}{2}G^{-1}(X_t)\nabla \log \pi(X_t)dt + \Lambda(X_t)dt + \sqrt{G^{-1}(X_t)}dB_t, \tag{43}$$

$$\Lambda_i(X_t) = \frac{1}{2}\sum_j \frac{\partial}{\partial x_j}\{G^{-1}(X_t)\}_{ij}. \tag{44}$$

It can be shown that this diffusion has invariant Lebesgue density $\pi(x)$, as required [33]. Intuitively, when a set is mapped onto the manifold, distances are changed by a factor, $\sqrt{G(x)}$. Therefore, to end up with the initial distances, they must first be changed by a factor of $\sqrt{G^{-1}(x)}$ before the mapping, which explains the volatility term in Equation (43).

The resulting Metropolis–Hastings proposal kernel for this "MALA on a manifold" was clarified in [33] and is given by:

$$Q(x,\cdot) \equiv \mathcal{N}\left(x + \frac{\lambda^2}{2}G^{-1}(x)\nabla \log \pi(x) + \lambda^2\Lambda(x), \lambda^2 G^{-1}(x)\right), \tag{45}$$

where λ^2 is a tuning parameter. The nonlinear drift term here is slightly different to that reported in [1,32], for reasons discussed in [33].

4.4. Choosing a Metric

We now turn to the question of which metric structure to put on the manifold, or equivalently, how to choose $G(x)$. In this section, we sometimes switch notation slightly, denoting the target density, $\pi(x|y)$, as some of the discussion is directed towards Bayesian inference, where $\pi(\cdot)$ is the posterior distribution for some parameter, x, after observing some data, y. The problem statement is: what is an appropriate choice of distance between points in the sample space of a given probability distribution?

A related (but distinct) question is how to define a distance between two probability distributions from the same parametric family, but with different parameters. This has been a key theme in information geometry, explored by Rao [46] and others [2] for many years. Although generic measures of distance between distributions (such as total variation) are often appropriate, based on information-theoretic principles, one can deduce that for a given parametric family, $\{p_x(y) : x \in \mathcal{X}\}$, it is in some sense natural to consider this "space of distributions" to be a manifold, where the Fisher information is the matrix, $G(x)$ (with the $\alpha = 0$ connection employed; see [2] for details).

Because of this, Girolami and Calderhead [1] proposed a variant of the Fisher metric for geometric Markov chain Monte Carlo, as:

$$G(x) = \mathbb{E}_{y|x}\left[-\frac{\partial^2}{\partial x_i \partial x_j}\log f(y|x)\right] - \frac{\partial^2}{\partial x_i \partial x_j}\log \pi_0(x), \tag{46}$$

where $\pi(x|y) \propto f(y|x)\pi_0(x)$ is the target density, f denotes the likelihood and π_0 the prior. The metric is tailored to Bayesian problems, which are a common use for MCMC, so the Fisher information is combined with the negative Hessian of the log-prior. One can also view this metric as the expected negative Hessian of the log target, since this naturally reduces to (46).

The motivation for a Hessian-style metric can also be understood from studying MCMC proposals. From (45) and by the same logic as for general pre-conditioning methods [32], the objective is to choose $G^{-1}(x)$ to match the covariance structure of $\pi(x|y)$ locally. If the target density were Gaussian with covariance matrix, Σ, then:

$$-\frac{\partial^2}{\partial x_i \partial x_j}\log \pi(x|y) = \Sigma. \tag{47}$$

In the non-Gaussian case, the negative Hessian is no longer constant, but we can imagine that it matches the correlation structure of $\pi(x|y)$ locally at least. Such ideas have been discussed in the geostatistics literature previously [47]. One problem with simply using (47) to define a metric is that unless $\pi(x|y)$ is log-concave, the negative Hessian will not be globally positive-definite, although Petra *et al.* [48] conjecture that it may be appropriate for use in some realistic scenarios and suggest some computationally efficient approximation procedures [48].

Example: Take $\pi(x) \propto 1/(1+x^2)$*, and set* $G(x) = -\partial^2 \log \pi(x)/\partial x^2$*. Then,* $G^{-1}(x) = (1+x^2)^2/(2-2x^2)$*, which is negative if* $x^2 > 1$*, so unusable as a proposal variance.*

Girolami and Calderhead [1] use the Fisher metric in part to counteract this problem. Taking expectations over the data ensures that the likelihood contribution to $G(x)$ in (46) will be positive (semi-)definite globally (e.g., [49]); so, provided a log-concave prior is chosen, then (46) should be a suitable choice for $G(x)$. Indeed, Girolami and Calderhead [1] provide several examples in which geometric MCMC methods using this Fisher metric perform better than their "non-geometric" counterparts.

Betancourt [50] also starts from the viewpoint that the Hessian (47) is an appropriate choice for $G(x)$ and defines a mapping from the set of $n \times n$ matrices to the set of positive-definite $n \times n$ matrices by taking a "smooth" absolute value of the eigenvalues of the Hessian. This is done in a way such that derivatives of $G(x)$ are still computable, inspiring the author to the name, SoftAbs metric. For a fixed value of x, the negative Hessian, $H(x)$, is first computed and, then, decomposed into $U^T D U$, where D is the diagonal matrix of eigenvalues. Each diagonal element of D is then altered by the mapping $t_\alpha : \mathbb{R} \to \mathbb{R}$, given by:

$$t_\alpha(\lambda_i) = \lambda_i \coth(\alpha \lambda_i), \tag{48}$$

where α is a tuning parameter (typically chosen to be as large as possible for which eigenvalues remain non-zero numerically). The function, t_α, acts as an absolute value function, but also uplifts eigenvalues, which are close to zero to $\approx 1/\alpha$. It should be noted that while the Fisher metric is only defined for models in which a likelihood is present and for which the expectation is tractable, the SoftAbs metric can be found for any target distribution, $\pi(\cdot)$.

Many authors (e.g., [1,48]) have noted that for many problems, the terms involving derivatives of $G(x)$ are often small, and so, it is not always worth the computational effort of evaluating them. Girolami and Calderhead [1] propose the simplified manifold, MALA, in which proposals are of the form:

$$Q(x,\cdot) \equiv \mathcal{N}\left(x + \frac{\lambda^2}{2}G^{-1}(x)\nabla \log \pi(x), \lambda^2 G^{-1}(x)\right) \tag{49}$$

Using this method means derivatives of $G(x)$ are no longer needed, so more pragmatic ways of regularising the Hessian are possible. One simple approach would be to take the absolute values of each eigenvalue, giving $G(x) = U^T|D|U$, where $H(x) = U^T DU$ is the negative Hessian and $|D|$ is a diagonal matrix with $\{|D|\}_{ii} = |\lambda_i|$ (this approach may fall into difficulties if eigenvalues are numerically zero). Another would be choosing $G(x)$ as the "nearest" positive-definite matrix to the negative Hessian, according to some distance metric on the set of $n \times n$ matrices. The problem has, in fact, been well-studied in mathematical finance, in the context of finding correlations using incomplete data sets [51], and tackled using distances induced by the Frobenius norm. Approximate solution algorithms are discussed in Higham [51]. It is not clear to us at present whether small changes to the Hessian would result in large changes to the corresponding positive definite matrix under a given distance or, indeed, whether given a distance metric on the space of matrices, there is always a well-defined unique "nearest" positive definite matrix. Below, we provide two simple examples, here showing how a "Hessian-style metric" can alleviate some of the difficulties associated with both heavy and light-tailed target densities.

Example: Take $\pi(x) \propto 1/(1+x^2)$*, and set* $G(x) = |-\partial^2 \log \pi(x)/\partial x^2|$*. Then,* $G^{-1}(x)\nabla \log \pi(x) = -x(1+x^2)/|1-x^2|$*, which no longer tends to zero as* $|x| \to \infty$*, suggesting a manifold variant of MALA with a Hessian-style metric may avoid some of the pitfalls of the standard algorithm. Note that the drift may become very large if* $|x| \approx 1$*, but since this event occurs with probability zero, we do not see it as a major cause for concern.*

Example: Take $\pi(x) \propto e^{-x^4}$*, and set* $G(x) = |-\partial^2 \log \pi(x)/\partial x^2|$*. Then,* $G^{-1}(x)\nabla \log \pi(x) = -x/3$*, which is* $O(x)$*, so alleviating the problem of spiralling proposals for light-tailed targets demonstrated by MALA in an earlier example.*

Other choices for $G(x)$ have been proposed, which are not based on the Hessian. These have the advantage that gradients need not be computed (either analytically or using computational methods). Sejdinovic *et al.* [52] propose a Metropolis–Hastings method, which can be viewed as a geometric variant of the RWM, where the choice for $G(x)$ is based on mapping samples to an appropriate feature space, and performing principal component analysis on the resulting features to choose a local covariance structure for proposals.

If we consider the RWM with Gaussian proposals to be a Euler–Maruyama discretisation of Brownian motion on a manifold, then proposals will take the form $Q(x,\cdot) \equiv \mathcal{N}(x + \lambda^2\Omega(x), \lambda^2 G^{-1}(x))$. If we assume (like in the simplified manifold MALA) that $\Omega(x) \approx 0$, then we have proposals centred at the current point in the Markov chain with a local covariance structure (the full Hastings acceptance rate must now be used as $q(x'|x) \neq q(x|x')$ in general).

As no gradient information is needed, the Sejdinovic *et al.* metric can be used in conjunction with the pseudo-marginal MCMC algorithm, so that $\pi(x|y)$ need not be known exactly. Examples from the article demonstrate the power of the approach [52].

An important property of any Riemannian metric is how it transforms under coordinate change (e.g., [2]). The Fisher information metric commonly studied in information geometry is an example of a "coordinate invariant" choice for $G(x)$. If we consider two parametrisations for a statistical model given

by x and $z = t(x)$, computing the Fisher information under x and then transforming this matrix using the Jacobian for the mapping, t, will give the same result as computing the Fisher information under z. It should be noted that because of either the prior contribution in (46) or the nonlinear transformations applied in other cases, none of the metrics we have reviewed here have this property, which means that we have no principled way of understanding how $G(x)$ will relate to $G(z)$. It is intuitive, however, that using information from all of $\pi(x)$, rather than only the likelihood contribution, $f(y|x)$, would seem sensible when trying to sample from $\pi(\cdot)$.

5. Survey of Applications

Rather than conduct our own simulation study, we instead highlight some cases in the literature where geometric MCMC methods have been used with success.

Martin *et al.* [53] consider Bayesian inference for a statistical inverse problem, in which a surface explosion causes seismic waves to travel down into the ground (the subsurface medium). Often, the properties of the subsurface vary with distance from ground level or because of obstacles in the medium, in which case, a fraction of the waves will scatter off these boundaries and be reflected back up to ground level at later times. The observations here are the initial explosion and the waves, which return to the surface, together with return times. The challenge is to infer the properties of the subsurface medium from this data. The authors construct a likelihood based on the wave equation for the data and perform Bayesian inference using a variant of the manifold MALA. Figures are provided showing the local correlations present in the posterior and, therefore, highlighting the need for an algorithm that can navigate the high density region efficiently. Several methods are compared in the paper, but the variant of MALA that incorporates a local correlation structure is shown to be the most efficient, particularly as the dimension of the problem increases [53].

Calderhead and Girolami [54] dealt with two models for biological phenomena based on nonlinear dynamical systems. A model of circadian control in the *Arabidopsis thaliana* plant comprised a system of six nonlinear differential equations, with twenty two parameters to be inferred. Another model for cell signalling consisted of a system of six nonlinear differential equations with eight parameters, with inference complicated by the fact that observations of the model are not recorded directly [54]. The resulting inference was performed using RWM, MALA and geometric methods, with the results highlighting the benefits of taking the latter approach. The simplified variant of MALA on a manifold is reported to have produced the most efficient inferences overall, in terms of effective sample size per unit of computational time.

Stathopoulos and Girolami [55] considered the problem of inferring parameters in Markov jump processes. In the paper, a linear noise approximation is shown, which can make inference in such models more straightforward, enabling an approximate likelihood to be computed. Models based on chemical reaction dynamics are considered; one such from chemical kinetics contained four unknown parameters; another from gene expression consisted of seven. Inference was performed using the RWM, the simplified manifold MALA and Hamiltonian methods, with the MALA reported as most efficient according to the chosen diagnostics. The authors note that the simplified manifold method is

both conceptually simple and able to account for local correlations, making it an attractive choice for inference [55].

Konukoglu *et al.* [56] designed a method for personalising a generic model for a physiological process to a specific patient, using clinical data. The personalisation took the form of patient-specific parameter inference. The authors highlight some of the difficulties of this task in general, including the complexity of the models and the relative sparsity of the datasets, which often result in a parameter identifiability issue [56]. The example discussed in the paper is the Eikonal-diffusion model describing electrical activity in cardiac tissue, which results in a likelihood for the data based on a nonlinear partial differential equation, combined with observation noise [56]. A method for inference was developed by first approximating the likelihood using a spectral representation and then using geometric MCMC methods on the resulting approximate posterior. The method was first evaluated on synthetic data and then repeated on clinical data taken from a study for ventricular tachycardia radio-frequency ablation [56].

6. Discussion

The geometric viewpoint in not necessary to understand manifold variants of the MALA. Indeed, several authors [32,33] have discussed these algorithms without considering them to be "geometric", rather simply Metropolis–Hastings methods in which proposal kernels have a position-dependent covariance structure. We do not claim that the geometric view is the only one that should be taken. Our goal is merely to point out that such position-dependent methods can often be viewed as methods defined on a manifold and that studying the structure of the manifold itself may lead to new insights on the methods. For example, taking the geometric viewpoint and noting the connection with information geometry enabled Girolami and Calderhead to adopt the Fisher metric for calculations [1]. We list here a few open questions on which the geometric viewpoint may help shed some insight.

Computationally-minded readers will have noted that using position-dependent covariance matrices adds a significant computational overhead in practice, with additional $O(n^3)$ matrix inversions required at each step of the corresponding Metropolis–Hastings algorithms. Clearly, there will be many problems for which the matrix, $G(x)$, does not change very much, and therefore, choosing a constant covariance $G^{-1}(x) = \Sigma$ may result in a more efficient algorithm overall. Geometrically, this would correspond to a manifold with scalar curvature close to zero everywhere. It may be that geometric ideas could be used to understand whether the manifold is flat enough that a constant choice of $G(x)$ is sufficient. To make sense of this truly would require a relationship between curvature, an inherently local property and more global statements about the manifold. Many results in differential geometry, beginning with the celebrated Gauss–Bonnet theorem, have previously related global and local properties in this way [57]. It is unknown to the authors whether results exist relating the curvature of a manifold to some global property, but this is an interesting avenue for further research.

A related question is when to choose the simplified manifold MALA over the full method. Problems in which the term, $\|\Lambda(x)\|$, is sufficient large to warrant calculation correspond to those for which the manifold has very high curvature in many places; so again, making some global statement related to curvature could help here.

Although there is a reasonably intuitive argument for why the Hessian is an appropriate starting point for $G(x)$, the lack of positive-definiteness may be seen as a cause for concern by some. After all, it could be argued that if the curvature is not positive-definite in a region, then how can it be a reasonable approximation to the local covariance structure. Many statistical models used to describe natural phenomena are characterised by distributions with heavy tails or multiple modes, for which this is the case. In addition, for target densities of the form $\pi(x) \propto e^{-|x|}$, the Hessian is everywhere equal to zero!The attempts to force positive-definiteness we have described will typically result in small moves being proposed in such regions of the sample space, which may not be an optimal strategy. Much work in information geometry has centred on the geometry of Hessian structures [58], and some insights from this field may help to better understand the question of choosing an appropriate metric. In addition, the field of pseudo-Riemannian geometry deals with forms of $G(x)$, which need not be positive-definite [39]; so again, understanding could be gained from here.

Some recent work in high-dimensional inference has centred on defining MCMC methods for which efficiency scales $O(1)$ with respect to the dimension, n, of $\pi(\cdot)$ [19,59]. In the case where X takes values in some infinite-dimensional function space, this can be done provided a Gaussian prior measure is defined for X. A striking result from infinite-dimensional probability spaces is that two different probability measures defined over some infinite dimensional space have a striking tendency to have disjoint supports [60]. The key challenge for MCMC is to define transition kernels for which proposed moves are inside the support for $\pi(\cdot)$. A straight-forward approach is to define proposals for which the prior is invariant, since the likelihood contribution to the posterior typically will not alter its support from that of the prior [19]. However, the posterior may still look very different from the prior, as noted in [61], so this proposal mechanism, though $O(1)$, can still result in slow exploration. Understanding the geometry of the support and defining methods that incorporate the likelihood term, but also respect this geometry, so as to ensure proposals remain in support of $\pi(\cdot)$, is an intriguing research proposition.

The methods reviewed in this paper are based on first order Langevin diffusions. Algorithms have also been developed that are based on second order Langevin diffusions, in which a stochastic differential equation governs the behaviour of the velocity of a process [62,63]. A natural extension to the work of Girolami and Calderhead [1] and Xifara *et al.* [33] would be to map such diffusions onto a manifold and derive Metropolis–Hastings proposal kernels based on the resulting dynamics. The resulting scheme would be a generalisation of [63], though the most appropriate discretisation scheme for a second order process to facilitate sampling is unclear and perhaps a question worthy of further exploration.

We have focused primarily here on the sample space $\mathcal{X} = \mathbb{R}^n$ and on defining an appropriate manifold on which to construct Markov chains. In some inference problems, however, the sample space is a pre-defined manifold, for example the set of $n \times n$ rotation matrices, commonly found in the field of directional statistics [64]. Such manifolds are often not globally mappable to Euclidean n-space. Methods have been devised for sampling from such spaces [65,66]. In order to use the methods described here for such problems, an appropriate approach for switching between coordinate patches at the relevant time would need to be devised, which could be an interesting area of further study.

Alongside these geometric problems, we can also discuss geometric MCMC methods from a statistical perspective. The last example given in the previous section hinted that the manifold MALA may cope better with target distributions with heavy tails. In fact, Latuszynski *et al.* [67] have shown

that, in one dimension, the manifold MALA is geometrically ergodic for a class of targets of the form $\pi(x) \propto \exp(-|x|^{\beta})$ for any choice of $\beta \neq 1$. This incorporates cases where tails are heavier than exponential and lighter than Gaussian, two scenarios under which geometric ergodicity fails for the MALA.

Finding optimal acceptance rates and scaling of λ with dimension are two other related challenges. In this case, the picture is more complex. Traditional results have been shown for Metropolis–Hastings methods in the case where target distributions are independent and identically-distributed or some other suitable symmetry and regularity in the shape of $\pi(\cdot)$. Manifold methods are, however, specifically tailored to scenarios in which this is not the case, scenarios in which there is a high correlation between components of x, which changes depending on the value of x. It is less clear how to proceed with finding relevant results that can serve as guidelines to practitioners here. Indeed, Sherlock [18] notes that a requirement for optimal acceptance rate results for the RWM to be appropriate is that the curvature of $\pi(x)$ does not change too much, yet this is the very scenario in which we would want to use a manifold method.

Acknowledgments

We thank the two reviewers for helpful comments and suggestions. Samuel Livingstone is funded by a PhD Scholarship from Xerox Research Centre Europe. Mark Girolami is funded by an Engineering and Physical Sciences Research Council Established Career Research Fellowship, EP/J016934/1, and a Royal Society Wolfson Research Merit Award.

Author Contributions

The article was written by Samuel Livingstone under the guidance of Mark Girolami. All authors have read and approved the final manuscript.

Appendix

A. Total Variation Distance

We show how to obtain (10) from (9). Denoting two probability distributions, $\mu(\cdot)$ and $\nu(\cdot)$, and associated densities, $\mu(x)$ and $\nu(x)$, we have:

$$\|\mu(\cdot) - \nu(\cdot)\|_{TV} := \sup_{A \in \mathcal{B}} |\mu(A) - \nu(A)|.$$

Define the set $B = \{x \in \mathcal{X} : \mu(x) > \nu(x)\}$. To see that $B \in \mathcal{B}$, note that $B = \cup_{q \in \mathbb{Q}} \{x \in \mathcal{X} : \mu(x) > q\} \cap \{x \in \mathcal{X} : \nu(x) < q\}$, and the result follows from properties of $\mathcal{B}$ (e.g., [68]). Now, for any $A \in \mathcal{B}$:

$$\mu(A) - \nu(A) \leq \mu(A \cap B) - \nu(A \cap B) \leq \mu(B) - \nu(B),$$

and similarly:

$$\nu(A) - \mu(A) \leq \nu(B^c) - \mu(B^c),$$

so, the supremum will be attained either at B or B^c. However, since $\mu(\mathcal{X}) = \nu(\mathcal{X}) = 1$, then:

$$[\mu(B) - \nu(B)] - [\nu(B^c) - \mu(B^c)] = 0,$$

so that

$$|\mu(B) - \nu(B)| = |\mu(B^c) - \nu(B^c)|.$$

Using these facts gives an alternative characterisation of the total variation distance as:

$$\begin{aligned}\|\mu(\cdot) - \nu(\cdot)\|_{TV} &= \frac{1}{2}\left(|\mu(B) - \nu(B)| + |\mu(B^c) - \nu(B^c)|\right) \\ &= \frac{1}{2}\int_{\mathcal{X}} |\mu(x) - \nu(x)|dx\end{aligned}$$

as required.

B. Gradient and Divergence Operators on a Riemannian Manifold

The gradient of a function on $\mathbb{R}^n$ is the unique vector field, such that, for any unit vector, u:

$$\langle \nabla f(x), u\rangle = D_u\left[f(x)\right] = \lim_{h\to 0}\left\{\frac{f(x+hu) - f(x)}{h}\right\}, \tag{50}$$

the directional derivative of f along u at $x \in \mathbb{R}^n$.

On a manifold, the gradient operator, ∇_M, can still be defined, such that the inner product $g_p(\nabla_M f(x), u) = D_u[f(x)]$. Setting $\nabla_M = G(x)^{-1}\nabla$ gives:

$$\begin{aligned}g_p(\nabla_M f(x), u) &= (G^{-1}(x)\nabla f(x))^T G(x)u, \\ &= \langle \nabla f(x), u\rangle,\end{aligned}$$

which is equal to the directional derivative along u as required.

The divergence of some vector field, v, at a point, $x \in \mathbb{R}^n$, is the net outward flow generated by v through some small neighbourhood of x. Mathematically, the divergence of $v(x) \in \mathbb{R}^3$ is given by $\sum_i \partial v_i/\partial x_i$. On a more general manifold, the divergence is also a sum of derivatives, but here, they are covariant derivatives. A short introduction is provided in Appendix C. Here, we simply state that the covariant derivative of a vector field, v, at a point $p \in M$ is the orthogonal projection of the directional derivative onto the tangent space, T_pM. Intuitively, a vector field on a manifold is a field of vectors, each of which lie in the tangent space to a point, $p \in M$. It only makes sense therefore to discuss how vector fields change along the manifold or in the direction of vectors, which also lie in the tangent space. Although the idea seems simple, the covariant derivative has some attractive geometric properties; notably, it can be completely written in local coordinates,and, so, does not depend on knowledge of an embedding in some ambient space.

The divergence of a vector field, v, defined on a manifold, M, at the point, $p \in M$, is defined as:

$$\text{div}_M(v) = \sum_{i=1}^{n} D^c_{e_i}[v_i],$$

where e_i denotes the i-th basis vector for the tangent space, T_pM, at $p \in M$, and v_i denotes the i-th coefficient. This can be written in local coordinates (see Appendix C) as:

$$\text{div}_M(v) = |G(x)|^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(|G(x)|^{\frac{1}{2}} v_i\right),$$

and can be combined with ∇_M to form the Laplace–Beltrami operator (41).

C. Vector Fields and the Covariant Derivative

Here, we provide a short introduction to vector fields and differentiation on a smooth manifold; see [38,39]. The following geometric notation is used here: (i) vector components are indexed with a superscript, e.g., $v = (v^1, ..., v^n)$; and (ii) repeated subscripts and superscripts are summed over, e.g., $v^i \mathbf{e}_i = \sum_i v^i \mathbf{e}_i$ (known as the Einstein summation convention).

For any smooth manifold, M, the set of all tangent vectors to points on M is known as the tangent bundle and denoted TM.

A C^r vector field defined on M is a mapping that assigns to each point, $p \in M$, a tangent vector, $v(p) \in T_pM$. In addition, the components of $v(p)$ in any basis for T_pM must also be C^r [38]. We will denote the set of all vector fields on M as $\Gamma(TM)$. For some vector field, $v \in \Gamma(TM)$, at any point, $p \in M$, the vector, $v(p) \in T_pM$, can be written as a linear combination of some n basis vectors $\{\mathbf{e}_1, ..., \mathbf{e}_n\}$ as $v = v^i \mathbf{e}_i$. To understand how v will change in a particular direction along M, it only makes sense, therefore, to consider derivatives along vectors in T_pM. Two other things must be considered when defining a derivative along a manifold: (i) how the components, v^i, of each basis vector will change; and (ii) how each basis vector, $\mathbf{e}_i$, itself will change. For the usual directional derivative on $\mathbb{R}^n$, the basis vectors do not change, as the tangent space is the same at each point, but for a more general manifold, this is no longer the case: the $\mathbf{e}_i$'s are referred to as a "local" basis for each T_pM.

The covariant derivative, D^c, is defined so as to account for these shortcomings. When considering differentiation along a vector, $u^* \notin T_pM$, u^* is simply projected onto the tangent space. The derivative with respect to any $u \in T_pM$ can now be decomposed into a linear combination of derivatives of basis vectors and vector components:

$$D^c_u[v] = D^c_{u^i \mathbf{e}_i}[v^i \mathbf{e}_i], \tag{51}$$

where the argument, p, has been dropped, but is implied for both components and local basis vectors. The operator, $D^c_u[v]$, is defined to be linear in both u and v and to satisfy the product rule [38]; so, Equation (51) can be decomposed into:

$$D^c_u[v] = u^i \left(D^c_{\mathbf{e}_i}[v^j] \mathbf{e}_j + v^j D^c_{\mathbf{e}_i}[\mathbf{e}_j] \right). \tag{52}$$

The operator, D^c, need, therefore, only be defined along the direction of basis vectors $\mathbf{e}_i$ and for vector component v^i and basis vector $\mathbf{e}_i$ arguments.

For components v^i, $D^c_{\mathbf{e_j}}[v^i]$ is defined as simply the partial derivative $\partial_j v^i := \partial v^i / \partial x^j$. The directional derivative of some basis vector $\mathbf{e}_i$ along some $\mathbf{e}_j$ is best understood through the example of a regular surface $\Sigma \subset \mathbb{R}^3$. Here, $D_{\mathbf{e}_j}[\mathbf{e}_i]$ will be a vector, $w \in \mathbb{R}^3$. Taking the basis for this space at the point, p, as $\{\mathbf{e}_1, \mathbf{e}_2, \hat{\mathbf{n}}\}$, where $\hat{\mathbf{n}}$ denotes the unit normal to $T_p\Sigma$, we can write $w = \alpha \mathbf{e}_1 + \beta \mathbf{e}_2 + \kappa \hat{\mathbf{n}}$. The covariant derivative, $D^c_{\mathbf{e}_j}[\mathbf{e}_i]$, is simply the projection of w onto $T_p\Sigma$, given by $w^* = \alpha \mathbf{e}_1 + \beta \mathbf{e}_2$. More generally, at some point, p, in a smooth manifold, M, the covariant derivative $D^c_{\mathbf{e}_j}[\mathbf{e}_i] = \Gamma^k_{ji} \mathbf{e}_k$ (with upper and lower indices summed over). The coefficients, Γ^k_{ji}, are known as the Christoffel symbols: Γ^k_{ji} denotes the coefficient of the k-th basis vector when taking the derivative of the i-th with respect to the j-th. If a Riemannian metric, g, is chosen for M; then, they can be expressed completely as a function

of g (or in local coordinates as a function of the matrix, G). Using these definitions, Equation (52) can be re-written as:

$$D_u^c[v] = u^i \left(\partial_i v^k + v^j \Gamma_{ij}^k\right) \mathbf{e}_k. \tag{53}$$

The divergence of a vector field, $v \in \Gamma(TM)$, at the point, $p \in M$, is given by:

$$\text{div}_M(v) = D_{\mathbf{e}_i}^c[v^i], \tag{54}$$

where, again, repeated indices are summed over. If $M = \mathbb{R}^n$, this reduces to the usual sum of partial derivatives, $\partial_i v^i$. On a more general manifold, M, the equivalent expression is:"'

$$D_{\mathbf{e}_i}^c[v^i] = \partial_i v^i + v^i \Gamma_{ij}^j, \tag{55}$$

where, again, repeated indices are summed. As has been previously stated, if a metric, g, and coordinate chart is chosen for M, the Christoffel symbols can be written in terms of the matrix, $G(x)$. In this case [69]:

$$\Gamma_{ij}^j = |G(x)|^{-\frac{1}{2}} \partial_i \left(|G(x)|^{\frac{1}{2}}\right), \tag{56}$$

so Equation (55) becomes:

$$D_{\mathbf{e}_i}^c[v^i] = |G(x)|^{-\frac{1}{2}} \partial_i \left(|G(x)|^{\frac{1}{2}} v^i\right), \tag{57}$$

where $v = v(x)$.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Girolami, M.; Calderhead, B. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Ser. B* **2011**, *73*, 123–214.
2. Amari, S.I.; Nagaoka, H. *Methods of Information Geometry*; American Mathematical Society: Providence, RI, USA, 2007; Volume 191.
3. Marriott, P.; Salmon, M. *Applications of Differential Geometry to Econometrics*; Cambridge University Press: Cambridge, UK, 2000.
4. Betancourt, M.; Girolami, M. Hamiltonian Monte Carlo for Hierarchical Models. **2013**, arXiv: 1312.0906.
5. Neal, R. MCMC using Hamiltonian Dynamics. In *Handbook of Markov Chain Monte Carlo*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2011; pp. 113–162.
6. Betancourt, M.; Stein, L.C. The Geometry of Hamiltonian Monte Carlo. **2011**, arXiv: 1112.4118.
7. Robert, C.P.; Casella, G. *Monte Carlo Statistical Methods*; Springer: New York, NY, USA, 2004; Volume 319.
8. Tierney, L. Markov chains for exploring posterior distributions. *Ann. Stat.* **1994**, *22*, 1701–1728.
9. Kipnis, C.; Varadhan, S. Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Commun. Math. Phys.* **1986**, *104*, 1–19.

10. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2012.
11. Plummer, M.; Best, N.; Cowles, K.; Vines, K. CODA: Convergence diagnosis and output analysis for MCMC. *R. News* **2006**, *6*, 7–11.
12. Gibbs, A.L.; Su, F.E. On choosing and bounding probability metrics. *Int. Stat. Rev.* **2002**, *70*, 419–435.
13. Jones, G.L.; Hobert, J.P. Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Stat. Sci.* **2001**, *16*, 312–334.
14. Jones, G.L. On the Markov chain central limit theorem. *Probab. Surv.* **2004**, *1*, 299–320.
15. Gelman, A.; Rubin, D.B. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **1992**, *7*, 457–472.
16. Sherlock, C.; Fearnhead, P.; Roberts, G.O. The random walk Metropolis: Linking theory and practice through a case study. *Stat. Sci.* **2010**, *25*, 172–190.
17. Sherlock, C.; Roberts, G. Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets. *Bernoulli* **2009**, *15*, 774–798.
18. Sherlock, C. Optimal scaling of the random walk Metropolis: General criteria for the 0.234 acceptance rule. *J. Appl. Probab.* **2013**, *50*, 1–15.
19. Beskos, A.; Kalogeropoulos, K.; Pazos, E. Advanced MCMC methods for sampling on diffusion pathspace. *Stoch. Processes Appl.* **2013**, *123*, 1415–1453.
20. Roberts, G.O.; Rosenthal, J.S. Optimal scaling for various Metropolis–Hastings algorithms. *Stat. Sci.* **2001**, *16*, 351–367.
21. Roberts, G.O.; Tweedie, R.L. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* **1996**, *83*, 95–110.
22. Mengersen, K.L.; Tweedie, R.L. Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Stat.* **1996**, *24*, 101–121.
23. Jarner, S.F.; Hansen, E. Geometric ergodicity of Metropolis algorithms. *Stoch. Processes Appl.* **2000**, *85*, 341–361.
24. Christensen, O.F.; Møller, J.; Waagepetersen, R.P. Geometric ergodicity of Metropolis–Hastings algorithms for conditional simulation in generalized linear mixed models. *Methodol. Comput. Appl. Probab.* **2001**, *3*, 309–327.
25. Neal, P.; Roberts, G. Optimal scaling for random walk Metropolis on spherically constrained target densities. *Methodol. Comput. Appl. Probab.* **2008**, *10*, 277–297.
26. Jarner, S.F.; Tweedie, R.L. Necessary conditions for geometric and polynomial ergodicity of random-walk-type. *Bernoulli* **2003**, *9*, 559–578.
27. Øksendal, B. *Stochastic Differential Equations*; Springer: New York, NY, USA, 2003.
28. Rogers, L.C.G.; Williams, D. *Diffusions, Markov Processes and Martingales: Volume 2, Itô Calculus*; Cambridge University Press: Cambridge, UK, 2000; Volume 2.

29. Meyn, S.P.; Tweedie, R.L. Stability of Markovian processes III: Foster–Lyapunov criteria for continuous-time processes. *Adv. Appl. Probab.* **1993**, *25*, 518–518.
30. Coffey, W.; Kalmykov, Y.P.; Waldron, J.T. *The Langevin Equation: with Applications to Stochastic Problems in Physics, Chemistry, and Electrical Engineering*; World Scientific: Singapore, Singapore, 2004; Volume 14.
31. Roberts, G.O.; Tweedie, R.L. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **1996**, *2*, 341–363.
32. Roberts, G.O.; Stramer, O. Langevin diffusions and Metropolis–Hastings algorithms. *Methodol. Comput. Appl. Probab.* **2002**, *4*, 337–357.
33. Xifara, T.; Sherlock, C.; Livingstone, S.; Byrne, S.; Girolami, M. Langevin diffusions and the Metropolis-adjusted Langevin algorithm. *Stat. Probab. Lett.* **2013**, *91*, 14–19.
34. Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. Ser. A Math. Phys. Sci.* **1946**, *186*, 453–461.
35. Critchley, F.; Marriott, P.; Salmon, M. Preferred point geometry and statistical manifolds. *Ann. Stat.* **1993**, *21*, 1197–1224.
36. Marriott, P. On the local geometry of mixture models. *Biometrika* **2002**, *89*, 77–93.
37. Barndorff-Nielsen, O.; Cox, D.; Reid, N. The role of differential geometry in statistical theory. *Int. Stat. Rev.* **1986**, *54*, 83–96.
38. Boothby, W.M. *An Introduction to Differentiable Manifolds and Riemannian Geometry*; Academic Press: San Diego, CA, USA, 1986; Volume 120.
39. Lee, J.M. *Smooth Manifolds*; Springer: New York, NY, USA, 2003.
40. Do Carmo, M.P. *Riemannian Geometry*; Springer: New York, NY, USA, 1992.
41. Nash, J.F., Jr. The imbedding problem for Riemannian manifolds. In *The Essential John Nash*; Princeton University Press: Princeton, NJ, USA, 2002; p. 151.
42. Manton, J.H. A Primer on Stochastic Differential Geometry for Signal Processing. **2013**, arXiv: 1302.0430.
43. Stewart, J. *Multivariable Calculus*; Cengage Learning: Boston, MA, USA, 2011.
44. Hsu, E.P. *Stochastic Analysis on Manifolds*; American Mathematical Society: Providence, RI, USA, 2002; Volume 38.
45. Kent, J. Time-reversible diffusions. *Adv. Appl. Probab.* **1978**, *10*, 819–835.
46. Radhakrishna Rao, C. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **1945**, *37*, 81–91.
47. Christensen, O.F.; Roberts, G.O.; Sköld, M. Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *J. Comput. Graph. Stat.* **2006**, *15*, 1–17.
48. Petra, N.; Martin, J.; Stadler, G.; Ghattas, O. A computational framework for infinite-dimensional Bayesian inverse problems: Part II. Stochastic Newton MCMC with application to ice sheet flow inverse problems. **2013**, arXiv: 1308.6221.
49. Pawitan, Y. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*; Oxford University Press: Oxford, UK, 2001.

50. Betancourt, M. A General Metric for Riemannian Manifold Hamiltonian Monte Carlo. In *Geometric Science of Information*; Springer: New York, NY, USA, 2013; pp. 327–334.
51. Higham, N.J. Computing the nearest correlation matrix—a problem from finance. *IMA J. Numer. Anal.* **2002**, *22*, 329–343.
52. Sejdinovic, D.; Garcia, M.L.; Strathmann, H.; Andrieu, C.; Gretton, A. Kernel Adaptive Metropolis–Hastings. **2013**, arXiv: 1307.5302.
53. Martin, J.; Wilcox, L.C.; Burstedde, C.; Ghattas, O. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM J. Sci. Comput.* **2012**, *34*, A1460–A1487.
54. Calderhead, B.; Girolami, M. Statistical analysis of nonlinear dynamical systems using differential geometric sampling methods. *Interface Focus* **2011**, *1*, 821–835.
55. Stathopoulos, V.; Girolami, M.A. Markov chain Monte Carlo inference for Markov jump processes via the linear noise approximation. *Philos. Trans. R. Soc. A* **2013**, *371*, 20110541.
56. Konukoglu, E.; Relan, J.; Cilingir, U.; Menze, B.H.; Chinchapatnam, P.; Jadidi, A.; Cochet, H.; Hocini, M.; Delingette, H.; Jaïs, P.; *et al.* Efficient probabilistic model personalization integrating uncertainty on data and parameters: Application to eikonal-diffusion models in cardiac electrophysiology. *Prog. Biophys. Mol. Biol.* **2011**, *107*, 134–146.
57. Do Carmo, M.P.; Do Carmo, M.P. *Differential Geometry of Curves and Surfaces*; Englewood Cliffs: Prentice-Hall, NJ, USA, 1976; Volume 2.
58. Shima, H. *The Geometry of Hessian Structures*; World Scientific: Singapore, Singapore, 2007; Volume 1.
59. Cotter, S.; Roberts, G.; Stuart, A.; White, D. MCMC methods for functions: Modifying old algorithms to make them faster. *Stat. Sci.* **2013**, *28*, 424–446.
60. Da Prato, G.; Zabczyk, J. *Stochastic Equations in Infinite Dimensions*; Cambridge University Press: Cambridge, UK, 2008.
61. Law, K.J. Proposals which speed up function-space MCMC. *J. Comput. Appl. Math.* **2014**, *262*, 127–138.
62. Ottobre, M.; Pillai, N.S.; Pinski, F.J.; Stuart, A.M. A Function Space HMC Algorithm With Second Order Langevin Diffusion Limit. **2013**, arXiv: 1308.0543.
63. Horowitz, A.M. A generalized guided Monte Carlo algorithm. *Phys. Lett. B* **1991**, *268*, 247–252.
64. Mardia, K.V.; Jupp, P.E. *Directional Statistics*; Wiley: New York, NY, USA, 2009; Volume 494.
65. Byrne, S.; Girolami, M. Geodesic Monte Carlo on embedded manifolds. *Scand. J. Stat.* **2013**, *40*, 825–845.
66. Diaconis, P.; Holmes, S.; Shahshahani, M. Sampling from a manifold. In *Advances in Modern Statistical Theory and Applications: A Festschrift in Honor of Morris L. Eaton*; Institute of Mathematical Statistics: Washington, DC, USA, 2013; pp. 102–125.
67. Latuszynski, K.; Roberts, G.O.; Thiery, A.; Wolny, K. Discussion on "Riemann manifold Langevin and Hamiltonian Monte Carlo methods" (by Girolami, M. and Calderhead, B.). *J. R. Stat. Soc. Ser. B* **2011**, *73*, 188–189.

68. Capinski, M.; Kopp, P.E. *Measure, Integral and Probability*; Springer: New York, NY, USA, 2004.
69. Schutz, B.F. *Geometrical Methods of Mathematical Physics*; Cambridge University Press: Cambridge, UK, 1984.