



# Likelihood-based Inference on Non- linear Spaces

**Using Diffusion Processes on  
Riemannian Manifolds**

M. A. Corstanje



# Likelihood-based Inference on Nonlinear Spaces

Using Diffusion Processes on Riemannian  
Manifolds

by

M. A. Corstanje

to obtain the degree of Master of Science  
at the Delft University of Technology,  
to be defended publicly on Wednesday August 28, 2019 at 14:00 AM.

Student number: 4303601  
Project duration: December 15, 2018 – August 28, 2019  
Thesis committee: Dr. ir. F. H. van der Meulen, TU Delft, supervisor  
Prof. dr. ir. M. C. Veraar, TU Delft  
Dr. J. Söhl, TU Delft

An electronic version of this thesis is available at <http://repository.tudelft.nl/>.



# Abstract

When data in higher dimensions with a certain constraint on it, say a set of locations on a sphere, is encountered, some classical statistical analysis methods fail, as the data no longer assumes its values in a linear space. In this thesis we consider such datasets and aim to do likelihood-based inference on the center of the data. To model the nonlinearity, we consider the data to be a set of points on a Riemannian manifold.

The general approach in this thesis comes from the classical result where the center can be represented as the maximum likelihood estimator for the true mean of the dataset. To model an underlying distribution we will model the data as observations of realizations of Brownian motion on the manifold observed at a fixed time and use the transition density of the Brownian motion to construct a likelihood. The likelihood can then be approximated using diffusion bridges.

This thesis thus first focuses on differential geometry as well as Itô and Stratonovich calculus. After that, we will introduce methods to construct a likelihood for the center of the dataset on a manifold before using simulated diffusion bridges to approximate this likelihood. We finish the thesis with some numerical experiments in Julia that demonstrate the results on the sphere.

**Keywords:** Diffusion processes, diffusion bridges, Riemannian manifolds, stochastic differential equations, stochastic simulation, geometric statistics.



# Acknowledgements

Before you lies the thesis, which is the result of nine months of hard work to obtain the title of Master of Science in Applied Mathematics at the TU Delft.

About 11 months ago, the first talks about this subject started with my supervisor, Frank van der Meulen. After visiting a seminar of the Applied Statistics group, we decided to start this research.

During the past months I have been researching this topic and in the process, I have been learning a lot in many different fields of mathematics, ranging from differential geometry to the simulation of diffusion bridges in the, to me very new, Julia programming language. After obtaining new knowledge in all these fields, it turned out to be possible to obtain new results and gain some very interesting insights of which I can be proud.

In May, my thesis took me to the university of Copenhagen to visit my supervisor and work with the research group in geometric statistics over there. I would like to thank the people there, Stefan Sommer in particular, for the time they took to improve this research.

I would also like to thank Mark Veraar and Jacob Söhl for taking the time to read this report and take place in my examination committee.

Last but not least would like to thank my Frank, my supervisor, for the many hours he took to Skype with me, read my intermediate results and work together in Copenhagen. His guidance and his work to accompany me throughout this research have been invaluable to me and I really enjoyed working together.

I hope that this thesis will be a pleasant reading!

*M. A. Corstanje  
Delft, August 2019*



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	General formulation of the problem	1
1.2	This approach on Euclidean spaces	2
1.3	Structure of the thesis	2
<b>2</b>	<b>Background Theory</b>	<b>3</b>
2.1	Riemannian manifolds	3
2.1.1	Differentiable manifolds and coordinate systems	3
2.1.2	Tangent spaces	4
2.1.3	Riemannian metrics	7
2.1.4	The Levi-Civita connection	8
2.1.5	Vector fields along curves, geodesics and the exponential map	9
2.1.6	Gradients, divergence and the Laplace-Beltrami operator	12
2.2	Stochastic processes: Itô and Stratonovich calculus	15
2.2.1	Stochastic processes	15
2.2.2	Stochastic integration and stochastic differential equations	17
2.2.3	Transition kernels, Kolmogorov equations and the infinitesimal generator	21
2.3	The Transition Density	23
2.3.1	Finding the transition density of a diffusion process	24
2.3.2	Parabolic equations	24
2.3.3	Conditions for the existence of a transition density	26
<b>3</b>	<b>Brownian Motion on a Riemannian manifold</b>	<b>27</b>
3.1	Brownian motion in local coordinates	27
3.1.1	Derivation of the SDE for Brownian motion in local coordinates	27
3.1.2	Discussion of this method	29
3.2	Regular Submanifolds of $\mathbb{R}^N$ and Projections	29
3.2.1	Regular submanifolds of $\mathbb{R}^N$	29
3.2.2	Projections and Brownian motion	30
3.2.3	Discussion of this method	33
<b>4</b>	<b>Stochastic Simulation of Diffusions and diffusion bridges</b>	<b>35</b>
4.1	The Euler-Maruyama method	35
4.2	Using 2-jets and the exponential map	38
4.3	Simulation of diffusion bridges using guided proposals	41
4.3.1	Choosing the auxiliary process	43
4.3.2	Numerical experiments of simulations of diffusion bridges	43
<b>5</b>	<b>Likelihood-Based Inference on Riemannian manifolds</b>	<b>45</b>
5.1	Approximating transition densities using diffusion bridges	45
5.2	Drawing samples from the likelihood of $\theta$	46
5.2.1	Uniform proposals	47
5.2.2	Langevin adjusted proposals	47
5.3	Maximum likelihood estimates	47
<b>6</b>	<b>Numerical experiments on the sphere</b>	<b>49</b>
6.1	Deriving the Itô SDE and calculations on a grid of points	49
6.2	Finding the maximum likelihood estimator	50
6.2.1	Stochastic gradient descent using the exponential map	51
6.2.2	Stochastic gradient descent while normalizing the updates	52

6.3	Markov Chain Monte Carlo methods	53
6.3.1	Uniform proposals	54
6.3.2	Langevin adjusted proposals	55
6.4	Notes	55
<b>7</b>	<b>Conclusion</b>	<b>57</b>
7.1	Characterizing Brownian motion on manifolds	57
7.2	Simulation of diffusions	57
7.3	Numerical experiments	57
7.4	Suggestions for future research	57
7.4.1	Extend the methods to other spaces	57
7.4.2	Weaken conditions for existence of transition densities	58
7.4.3	Other approximations of the the transition density $p$	58
7.4.4	Relevance of the parameter $T$	58
7.4.5	Choices for the auxiliary process	58
	<b>Bibliography</b>	<b>59</b>
<b>A</b>	<b>Additional Theory on Martingales and stopping times</b>	<b>61</b>
A.1	Conditional expectation	61
A.2	Martingales and stopping times	62
A.3	Quadratic (co)variation	62
<b>B</b>	<b>Markov Chain Monte Carlo methods and the Metropolis-Hastings algorithm</b>	<b>65</b>
B.1	Discrete time Markov chains	65
B.2	Markov chain Monte Carlo methods	66
B.2.1	Langevin adjusted propals	66
	<b>Index</b>	<b>69</b>

# Introduction

In many statistical applications, we encounter data that takes values in more than one dimension but with certain constraints. One can think for example of a set of locations on earth, where the data takes values in three dimensions but has the constraint of lying on a sphere. Doing statistical inference when data takes its values in a nonlinear space is quite different from the regular setting where many common notions in statistics, such as the mean or standard deviation of a set of data, have been researched for a long time. They are usually easy to calculate in linear spaces since the concept of adding vectors and multiplication by constants is present. As an example, consider a set of points in a plane. The mean of these points will simply be the point that lies in the middle and is found by adding all the vectors and scaling the result with the sample size. However, if all these points were to lie on a circle, then the mean that is calculated is not necessarily on the circle, because the circle is not a linear space and thus a linear combination of vectors that take their values on it does not have to stay on the circle. The aim of this thesis is to set up a model for performing statistical inference on the center of a dataset when data takes values in a nonlinear space.

## 1.1. General formulation of the problem

In this thesis, we will consider datasets consisting of random elements  $\xi_1, \dots, \xi_n$  of a  $d$ -dimensional Riemannian manifold  $(\mathcal{M}, g)$ . A toy example will be the sphere, or in some cases the circle or the torus. The aim of this thesis is to do inference on the center of the dataset represented by a parameter  $\theta$ .

In the case of a dataset modeled as an independent, identically distributed  $\mathcal{N}(\theta, \sigma^2)$  sample, the center of a dataset, the sample mean, coincides with the maximum likelihood estimator for the true mean  $\theta$  of the underlying distribution. When a sample has values on a nonlinear space, however, the notion of an underlying distribution becomes quite abstract. For example, there is no straightforward way of characterizing normal distributions on the sphere. The first alternative for a center of a dataset is the Fréchet mean  $\operatorname{argmin}_{\mu} \sum_{i=1}^n d(\xi_i, \mu)^2$  where  $d$  is a metric on  $\mathcal{M}$ . This alternative, however, does not give us a sense of uncertainty or an idea of the distribution of the center and there is also no straightforward way of selecting the metric.

In this thesis, we propose to view the center as the most likely starting point for a set of  $n$  realizations of Brownian motion on  $\mathcal{M}$  that hit the data points at a given time  $T$ . We will thus use the transition density of Brownian motion to characterize a distribution for the center of the sample. Mathematically, we can formulate this as follows.

Let  $X = (X_t)_t$  be a Brownian motion on  $\mathcal{M}$  starting at  $\theta$  and  $T > 0$  be a fixed end time. Furthermore, let  $(x_t^i)_t, i = 1, \dots, n$  denote  $n$  independent sample paths of the process  $X$ . We aim to find the distribution of  $\theta \mid x_T^1 = \xi_1, \dots, x_T^n = \xi_n$ , which can be interpreted as a distribution for the center of the data.

## 1.2. This approach on Euclidean spaces

If we use this approach in Euclidean spaces, we consider the dataset to be realizations of independent observations  $X_T^{(1)}, \dots, X_T^{(n)}$  of standard Brownian motions starting at  $\theta$ . The data points are thus observations of the diffusions governed by the stochastic differential equations

$$dX_t^{(i)} = dW_t^{(i)} \quad X_0^{(i)} = \theta, \quad i = 1, \dots, n$$

The likelihood function for  $\theta$  is given by the product of the transition densities of the diffusions.

$$f(\theta) = \prod_{i=1}^n p(0, \theta; T, X_T^{(i)})$$

where  $p$  is such that  $\mathbb{P}(X_t \in dy \mid X_s = x) = p(s, x; t, y) dy$ . Since a standard Brownian motion in a Euclidean space has a normal transition density, the maximum likelihood estimator for  $\theta$  coincides with the classical sample mean.

This result motivates the approach of this thesis on a Riemannian manifold since methods for describing Brownian motion on manifolds have been derived in previous works and can now be applied to this research.

## 1.3. Structure of the thesis

In order to get to the goal formulated in the previous section, we will have to go through various steps. The thesis starts off with some background theory. This includes an introduction to Riemannian geometry, preliminaries and basic definitions of probabilistic notions and a couple of results regarding the transition density of a Markov process. In this chapter we first introduce Riemannian manifolds in various steps and derive various concepts in Riemannian geometry such as tangent spaces and the Laplace-Beltrami operator. After that, we introduce both Itô and Stratonovich stochastic calculus and formulate some theorems and definitions in that field before delving a bit deeper into transition densities of stochastic process, as these play an important part in the likelihood. In chapter 2.3, we delve deeper into the transition density of Brownian motion on a manifold via the Kolmogorov equations. In a slight intermezzo to functional analysis in parabolic equations, we will describe conditions on when we can solve Kolmogorov's equations.

In chapter 3, we discuss Brownian motion on Riemannian manifolds. We first introduce a characterization of Brownian motion in local coordinates via the Laplace-Beltrami operator in chapter 3.1 and after that we move to a characterization via orthogonal projections in chapter 3.2.

Chapter 4 moves away from the theoretic background and towards stochastic simulations. By successfully simulating diffusion bridges on a manifold, we can approximate the log-likelihood of  $\theta$ . For this reason, we discuss common algorithms used to simulate diffusions. Here, we discuss both the Euler-Maruyama method and jets or functions like the exponential map when known in closed form. In this chapter, we also discuss the simulation of diffusion bridges using guided proposals.

In chapter 5, we derive a method to use the simulations of diffusion bridges to approximate the likelihood of the center of the data and discuss various methods to draw samples from the likelihood and obtain maximum likelihood estimates. We finish the thesis by testing this method on the sphere by applying various algorithms to determine maximum likelihood estimates on simulated data and draw samples from the likelihood.

# 2

## Background Theory

The background theory in this chapter includes various widely used notions in the field of stochastic processes and Riemannian geometry.

### 2.1. Riemannian manifolds

In this section, we recap some results from differential geometry in order to familiarize the reader with Riemannian manifolds and introduce various frequently used notation. The definitions and notations used in this chapter are based on Spivak, 1970 [25], Lovett, 2010 [14] and Hsu, 2002 [9].

#### 2.1.1. Differentiable manifolds and coordinate systems

The theory of differential geometry is widely used in order to understand subspaces of  $\mathbb{R}^N$  that are no longer linear, but do, locally, have a certain similarity to  $\mathbb{R}^d$  for some  $d < N$ .

**Definition 2.1.1** (Manifold). *A manifold  $\mathcal{M}$  of dimension  $d$  is a Hausdorff topological space such that for each point  $p \in \mathcal{M}$ , there exists an open neighborhood  $U$  of  $p$  in  $\mathcal{M}$  and a homeomorphism  $x : U \rightarrow \mathbb{R}^d$ .*

From definition 2.1.1 is not immediately obvious that there is a structure on arbitrary manifolds as elements of the manifold can have multiple neighborhoods that can be mapped homeomorphically to completely different sets in  $\mathbb{R}^d$ . In order to do analysis on a manifold, we need to add structure to it.

**Definition 2.1.2** (Atlas). *Let  $\mathcal{M}$  be a  $d$ -dimensional manifold. An atlas  $\mathcal{A}$  is a set of homeomorphisms  $x_\alpha : U_\alpha \rightarrow \mathbb{R}^d$  such that  $\bigcup_\alpha U_\alpha = \mathcal{M}$  and for all  $\alpha$  and  $\beta$ , the transition functions*

$$\begin{aligned}x_\alpha \circ x_\beta^{-1} &: x_\beta(U_\alpha \cap U_\beta) \rightarrow x_\alpha(U_\alpha \cap U_\beta) \\x_\beta \circ x_\alpha^{-1} &: x_\alpha(U_\alpha \cap U_\beta) \rightarrow x_\beta(U_\alpha \cap U_\beta)\end{aligned}$$

*are functions between subsets of  $\mathbb{R}^d$  of class  $\mathcal{C}^1$ . A manifold  $\mathcal{M}$  endowed with an atlas  $\mathcal{A}$  is called a differentiable manifold and the elements  $(x_\alpha, U_\alpha)$  of  $\mathcal{A}$  are called charts or (local) coordinate systems. If the transition functions are of class  $\mathcal{C}^\infty$ , we speak of a smooth manifold.*

In some literature (see e.g. Lovett, 2010 [14]), manifolds are even defined via atlases, as these guarantee that calculations can be done nicely on them. Smoothness of the transition functions ensures that, when walking on the manifold, the switches between charts happen smoothly and therefore we can define various concepts from analysis via local behavior.

An elementary example of a smooth manifolds is  $\mathbb{R}^d$  itself with just one chart being the identity map. One can also endow the sphere  $\mathbb{S}^2 = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\}$  in  $\mathbb{R}^3$  with various atlases, such as the projections from the points  $(0, 0, 1)$  and  $(0, 0, -1)$  onto  $\mathbb{R}^2$ , turning it into a smooth manifold (see Spivak, 1970 [25]) of dimension 2, or consider the spherical coordinate system  $(u, v)$ , which has inverse

$$\begin{aligned}x &= \cos u \sin v \\y &= \sin u \sin v \\z &= \cos v\end{aligned}$$

for  $(u, v) \in [0, 2\pi] \times [0, \pi]$ . Obviously, the same can be done for the circle  $S^1 = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$  with the parameterization  $u \mapsto (\cos u, \sin u)$ .

### 2.1.2. Tangent spaces

A very important notion in the field of differentiable manifolds is the notion of a tangent space, but in order to understand this notion, we first need to know about (directional) derivatives. Since manifolds locally behave in a similar way to  $\mathbb{R}^d$ , notions such as differentiable functions can be defined via charts.

**Definition 2.1.3** (Differentiable functions between manifolds). *A function  $f : \mathcal{M} \rightarrow \mathcal{N}$  between two smooth manifolds of dimensions  $d$  and  $d'$  is differentiable if the map  $y \circ f \circ x^{-1} : x(U \cap f^{-1}(V)) \rightarrow y(V)$  is a differentiable map between subsets of  $\mathbb{R}^d$  and  $\mathbb{R}^{d'}$  for all charts  $(x, U)$  in  $\mathcal{M}$  and  $(y, V)$  in  $\mathcal{N}$ .*

**Example 2.1.4** (Differentiable curves on manifolds). *Using this definition, we can now characterize differentiable curves on a manifold. Let  $\gamma : \mathbb{R} \rightarrow \mathcal{M}$  be a curve on a differentiable manifold  $\mathcal{M}$ . Note that  $\mathbb{R}$  is a differentiable manifold with just one chart given by  $(\text{id}, \mathbb{R})$ . A curve is thus differentiable if the map  $\mathbb{R} \cap \gamma^{-1}(U) \rightarrow x(U)$  given by  $t \mapsto x(\gamma(t))$  is a differentiable map between subsets of  $\mathbb{R}$  for all charts  $(x, U)$  in  $\mathcal{M}$ .*

Knowing this, we can also define directional derivatives.

**Definition 2.1.5** (Directional derivative). *Let  $\mathcal{M}$  be a  $d$ -dimensional differentiable manifold,  $p \in \mathcal{M}$  and let  $U \subseteq \mathcal{M}$  be an open neighborhood of  $p$  in  $\mathcal{M}$ . Furthermore, let  $\varepsilon > 0$  and let  $\gamma : (-\varepsilon, \varepsilon) \rightarrow \mathcal{M}$  be a differentiable curve such that  $\gamma(0) = p$ . For any differentiable function  $f : U \rightarrow \mathbb{R}$ , we define the directional derivative of  $f$  along  $\gamma$  as*

$$D_\gamma f := \left. \frac{d}{dt} f(\gamma(t)) \right|_{t=0} \quad (2.1)$$

The operator  $D_\gamma$  is called the tangent vector to  $\gamma$  at  $p$ .

These tangent vectors will help us in defining the tangent space to  $\mathcal{M}$  at the point  $p$ , but in order to give a proper definition, we first need the following proposition

**Proposition 2.1.6.** *Let  $\mathcal{M}$  be a  $d$ -dimensional differentiable manifold and let  $p \in \mathcal{M}$ . The set of all directional derivatives at  $p$  is a vector space of dimension  $d$  with basis  $\left\{ \frac{\partial}{\partial x^1}, \dots, \frac{\partial}{\partial x^d} \right\}$*

*Proof [Adapted from Lovett, 2010 [14]. ]* Let  $p \in \mathcal{M}$  and let  $x : U \rightarrow \mathbb{R}^d$  be a coordinate system around  $p$ . Now define the line curve  $v_i : (-\varepsilon, \varepsilon) \rightarrow \mathcal{M}$  via  $v_i(t) = x^{-1}(0, \dots, 0, t, 0, \dots, 0)$ , where the  $t$  appears in the  $i$ th place. Then

$$D_{v_i} f = \left. \frac{d}{dt} f(x^{-1}(0, \dots, 0, t, 0, \dots, 0)) \right|_{t=0} = \left. \frac{\partial f}{\partial x^i}(x^{-1}(0, \dots, t, \dots, 0)) \right|_{t=0} = \left. \frac{\partial f}{\partial x^i} \right|_p$$

We thus see that  $D_{v_i} = \frac{\partial}{\partial x^i} \Big|_p$ . Now for any differentiable curve  $\gamma : (-\varepsilon, \varepsilon) \rightarrow \mathcal{M}$ , we can write  $\gamma$  in local coordinates as  $x \circ \gamma(t) = (\gamma^1(t), \dots, \gamma^d(t))$ , where  $\gamma^i(t) = x^i(\gamma(t))$ . This gives us

$$\begin{aligned} D_\gamma f &= \left. \frac{d}{dt} f \circ x^{-1}(\gamma^1(t), \dots, \gamma^d(t)) \right|_{t=0} \\ &= \sum_{i=1}^d \left. \frac{\partial f}{\partial x^i} \right|_p \left. \frac{d\gamma^i}{dt} \right|_{t=0} \end{aligned}$$

and hence  $D_\gamma f$  is a linear combination of the operators  $\frac{\partial}{\partial x^i} \Big|_p$ . Since these operators are also independent, they form a basis for the space of tangent vectors at  $p$  and this space must therefore have dimension  $d$ .  $\square$

**Definition 2.1.7** (Tangent space and tangent bundle). *The space of all tangent vectors at a point  $p \in \mathcal{M}$  is called the tangent space at  $p$  and will be denoted as  $T_p \mathcal{M}$ . The disjoint union of all tangent spaces is the tangent bundle and denoted by  $T\mathcal{M}$ . Elements of  $T\mathcal{M}$  are given by tuples  $(p, X_p)$  where  $p \in \mathcal{M}$  and  $X_p \in T_p \mathcal{M}$ .*

Tangent spaces enable us to define differentials as well.

**Definition 2.1.8** (Differential). *Let  $\mathcal{M}$  and  $\mathcal{N}$  be differentiable manifolds and let  $\phi : \mathcal{M} \rightarrow \mathcal{N}$  be a differentiable map. The differential of  $\phi$  at  $p \in \mathcal{M}$  is defined as the map  $d\phi_p : T_p\mathcal{M} \rightarrow T_{\phi(p)}\mathcal{N}$  given by  $D_\gamma \mapsto D_{\phi \circ \gamma}$*

**Example 2.1.9** (Tangent space to the circle). *Let  $p = (\cos u, \sin u)$  be a point on the circle and let  $\gamma(t) = (\cos u(t), \sin u(t))$  be a curve with  $\gamma(0) = p$ . Then  $\gamma'(t) = (-\sin u(t), \cos u(t)) \frac{du}{dt}$ . Hence*

$$D_\gamma = \begin{pmatrix} -\sin u(0) \\ \cos u(0) \end{pmatrix} u'(0) \frac{\partial}{\partial u} \Big|_p$$

*Now let us interpret this result. Note that the vector  $(-\sin u(0), \cos u(0))$  is orthogonal to  $p$  and  $u'(0)$  a constant is of dimension 1. The tangent space to the circle at  $p$  is therefore a line that is orthogonal to  $p$ . Note that this line is tangent to the circle and thus matches the elementary geometrical interpretation of a tangent space to a circle.*

**Example 2.1.10** (Tangent space to the sphere). *We will apply the same method to the sphere at the point  $p = (\cos u \sin v, \sin u \sin v, \cos v)$  and a curve  $\gamma(t) = (\cos u(t) \sin v(t), \sin u(t) \sin v(t), \cos v(t))$  with  $\gamma(0) = p$ . Standard calculations show us that*

$$\frac{d\gamma}{dt} = \begin{pmatrix} -\sin u(t) \sin v(t) \\ \cos u(t) \sin v(t) \\ 0 \end{pmatrix} \frac{du}{dt} + \begin{pmatrix} \cos u(t) \cos v(t) \\ \sin u(t) \cos v(t) \\ -\sin v(t) \end{pmatrix} \frac{dv}{dt}$$

*Note that these vectors are both orthogonal to  $p$  at  $t = 0$  and are orthogonal to each other as well. They will thus span the plane that is orthogonal to  $p$ , which is thus the tangent space to the sphere at  $p$ .*

In the definition of tangent spaces 2.1.7 we introduced the tangent bundle  $T\mathcal{M}$  that is given by the disjoint union of all tangent spaces to  $\mathcal{M}$ . We can now consider the canonical projection map  $\pi : T\mathcal{M} \rightarrow \mathcal{M}$  mapping  $(p, X_p) \in T\mathcal{M}$  to  $p \in \mathcal{M}$ . For any  $p \in \mathcal{M}$  we then have that  $\pi^{-1}(\{p\}) = T_p\mathcal{M}$ , which we call the fiber of  $p$ . Furthermore if  $A \subseteq \mathcal{M}$ , we call a map  $\sigma : A \rightarrow T\mathcal{M}$  such that  $\pi \circ \sigma = \text{id}$  a section of  $T\mathcal{M}$  in  $A$ . Notice that if  $\mathcal{M}$  is a smooth manifold, that the tangent spaces transform smoothly into each other as well. This process of associating the manifold with a vector space at each point generalizes to the concept of a vector bundle.

**Definition 2.1.11** (Vector bundles). *Let  $\mathcal{M}$  be a differentiable manifold of dimension  $d$  with atlas  $\mathcal{A} = \{(\phi_\alpha, U_\alpha)\}_{\alpha \in I}$  and let  $V$  be a finite-dimensional, real vector space. A vector bundle over  $\mathcal{M}$  of fiber  $V$  is a Hausdorff topological space  $E$  with a continuous surjection  $\pi : E \rightarrow \mathcal{M}$  (the bundle projection) and a collection  $\Psi$  of homeomorphisms (trivializations)  $\psi_\alpha : U_\alpha \times V \rightarrow \pi^{-1}(U_\alpha)$  such that if  $U_\alpha \cap U_\beta \neq \emptyset$ , then*

$$\psi_\beta^{-1} \circ \psi_\alpha : (U_\alpha \cap U_\beta) \times V \rightarrow (U_\alpha \cap U_\beta) \times V$$

*is of the form*

$$\psi_\beta^{-1} \circ \psi_\alpha(p, v) = (p, \theta_{\beta\alpha}(p)v)$$

*where  $\theta_{\beta\alpha}(p) : U_\alpha \cap U_\beta \rightarrow \text{GL}(V)$  is a continuous map into the general linear group.*

A vector bundle is often denoted by  $\xi$ . The space  $E$  is called the total space and denoted  $E(\xi)$  while the manifold  $\mathcal{M}$  is called the base space and denoted  $B(\xi)$ .

The idea behind a vector bundle is to associate a vector space  $V$  that "connects" the manifold to the bundle locally homeomorphically. For any chart  $U_\alpha$ , there is a homeomorphism  $\psi_\alpha$  that maps  $U_\alpha \times V$  homeomorphically into a subspace  $\pi^{-1}(U_\alpha)$  of the total space  $E(\xi)$ . This is not where it stops however, as the homeomorphisms also transition into one another in a continuous way through the transitions  $\theta_{\beta\alpha}$ .

Some trivial examples of vector bundles are the topological space  $\mathcal{M} \times V$  for any real vector space  $V$  where the trivialization maps are just the identity maps on  $U_\alpha \times V$  and the maps  $\theta_{\beta\alpha}$  are the identity linear transformation maps. Another example is the tangent bundle  $T\mathcal{M}$  that we hinted at earlier in this section.  $T\mathcal{M}$  is a vector bundle over  $\mathcal{M}$  with fiber  $\mathbb{R}^d$ , see proposition 4.1.4 of Lovett, 2010 [14] for the proof.

**Definition 2.1.12** (Global section). *Let  $\xi$  be a vector bundle over a manifold  $\mathcal{M}$  of fiber  $V$  with projection  $\pi : E(\xi) \rightarrow \mathcal{M}$ . A global section of  $\xi$  is a continuous map  $s : \mathcal{M} \rightarrow E(\xi)$  such that  $\pi \circ s$  is the identity function on  $\mathcal{M}$ . The set of all global sections is denoted by  $\Gamma(\xi)$ .*

To illustrate the intuition behind a global section, we go back to the example of the tangent bundle. Observe that a global section is the map that assigns to each point  $p \in \mathcal{M}$  the tuple  $(p, X_p)$ , where  $X_p$  has to be an element of  $T_p\mathcal{M}$ . It does not matter to which vector  $X_p \in T_p\mathcal{M}$  the point  $p$  is mapped for this mapping to be a global section as the canonical projection  $\pi$  maps all these tuples back to  $p$ . A global section of  $T\mathcal{M}$  is thus a continuous map that assigns a vector in the tangent space to each point. When we consider a global section of  $T\mathcal{M}$  we will refer to it as a *vector field*. The set of all vector fields on  $\mathcal{M}$  is denoted by  $\Gamma(T\mathcal{M})$ . A vector field  $X$  is of class  $\mathcal{C}^k$  if  $X : \mathcal{M} \rightarrow T\mathcal{M}$  is a map of class  $\mathcal{C}^k$  between manifolds and we typically talk about smooth vector fields when the map is of class  $\mathcal{C}^\infty$ . Figure 2.1 demonstrates this by showing an example of a vector field on the sphere.

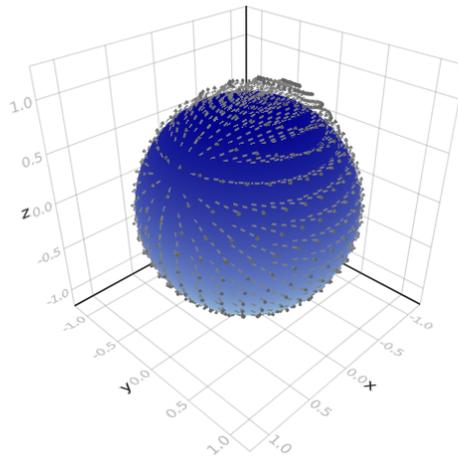


Figure 2.1: An illustration of an arbitrary vector field on the sphere where each arrow denotes a vector in the tangent space to a point.

**Proposition 2.1.13.** *Let  $X$  and  $Y$  be two vector fields of class  $\mathcal{C}^1$  on a differentiable manifold  $\mathcal{M}$ . Then the operation  $f \mapsto (XY - YX)f$  is another vector field.*

*Proof.* See proposition 4.2.9 of Lovett, 2010 [14]. □

**Definition 2.1.14** (Lie bracket). *Let  $X$  and  $Y$  be two vector fields of class  $\mathcal{C}^1$  on a differentiable manifold  $\mathcal{M}$ . The vector field  $[X, Y] := XY - YX$  is called the Lie bracket of  $X$  and  $Y$ . Notice that, since  $X$  and  $Y$  act on functions, we have  $[X, Y](f) = X(Y(f)) - Y(X(f))$ .*

For all  $a, b \in \mathbb{R}$  and differentiable functions  $f, g : \mathcal{M} \rightarrow \mathbb{R}$ , the following properties hold

1. Anticommutativity:  $[X, Y] = -[Y, X]$ .
2. Bilinearity:  $[aX + bY, Z] = a[X, Y] + b[Y, Z]$  and similarly for the second input variable.
3.  $[[X, Y], Z] + [[Y, Z], X] + [[Z, X], Y] = 0$ , this is called the Jacobi identity.
4.  $[fX, gY] = fg[X, Y] + fX(g)Y - gY(f)X$ .

### 2.1.3. Riemannian metrics

Up to this point, we have considered manifolds to be sets that are locally alike  $\mathbb{R}^d$  with a differentiable structure. We improve on this structure by introducing Riemannian metrics. Recall from proposition 2.1.6 that the tangent space to a point on a manifold is a vector space of dimension  $d$ . This motivates the following definition

**Definition 2.1.15** (Riemannian metric). *Let  $\mathcal{M}$  be a smooth manifold. A Riemannian metric on  $\mathcal{M}$  is a family of (positive definite) inner products*

$$g_p : T_p\mathcal{M} \times T_p\mathcal{M} \rightarrow \mathbb{R}, \quad p \in \mathcal{M}$$

*such that, for every pair of differentiable vector fields  $X, Y$  on  $\mathcal{M}$ , the map  $p \mapsto g_p(X_p, Y_p)$  defines a smooth function  $\mathcal{M} \rightarrow \mathbb{R}$ . A smooth manifold equipped with a Riemannian metric is referred to as a Riemannian manifold and denoted  $(\mathcal{M}, g)$ .*

Since, at a given  $p \in \mathcal{M}$ , the Riemannian metric  $g$  gives us an inner product on the space  $T_p\mathcal{M}$ , which has basis  $\left\{ \frac{\partial}{\partial x^1}, \dots, \frac{\partial}{\partial x^d} \right\}$ , by proposition 2.1.6, we can represent the metric at  $p$  by the matrix  $(g_{ij})$  with entries  $g_{ij} = g_p\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right)$ , which is also frequently used notation for Riemannian metrics. Note that with this new notation, one can write the Riemannian metric applied to  $X = \sum_{i=1}^d X^i \frac{\partial}{\partial x^i} \in T_p\mathcal{M}$  and  $Y = \sum_{i=1}^d Y^i \frac{\partial}{\partial x^i} \in T_p\mathcal{M}$  in local coordinates as

$$g_p(X, Y) = \sum_{i=1}^d \sum_{j=1}^d g_{ij} X^i Y^j \quad (2.2)$$

For a fixed positive definite matrix  $(g_{ij})$ . We will denote the inverse matrix in a similar way:  $(g^{ij}) = (g_{ij})^{-1}$ . To reduce the complexity of most equations, we use the following convention from now on.

**Definition 2.1.16** (Einstein's summation convention). *In Einstein notation, any index that appears both in sub- and in superscript is assumed to be summed over all of its possible values.*

In Einstein notation, equation (2.2) can thus be briefly written as  $g_p(X, Y) = g_{ij} X^i Y^j$ . To simplify the notation in many equations, we will abbreviate  $\frac{\partial}{\partial x^i}$  by  $\partial_i$  from now on.

**Example 2.1.17.** *A trivial Riemannian metric emerges when a subset  $\mathcal{M} \subseteq \mathbb{R}^N$  forms a manifold of dimension  $d < N$ . This metric is found by simply taking the dot product in  $\mathbb{R}^N$ . That is, for  $p \in \mathcal{M}$ , define  $g_p$  to be the bilinear map  $g_p : T_p\mathcal{M} \times T_p\mathcal{M} \rightarrow \mathbb{R}$  given by  $g_p(u, v) = u \cdot v$ . In local coordinates  $(x^1, \dots, x^d)$ , the corresponding matrix is given by the  $d \times d$  matrix*

$$(g_{ij}) = (\partial_i \cdot \partial_j).$$

*In this case, we say that  $\mathcal{M}$  inherits the Riemannian metric from  $\mathbb{R}^N$ .*

**Example 2.1.18** (The circle  $\mathbb{S}^1$ ). *The circle can be parameterized via the map  $x : [0, 2\pi] \rightarrow \mathbb{S}^1$  given by  $x(u) = (\cos u, \sin u)$  where  $u \in [0, 2\pi]$ . Since this manifold is embedded in  $\mathbb{R}^2$ , we can inherit the standard inner product on  $\mathbb{R}^2$ .  $\mathbb{S}^1$  is a manifold of dimension 1 and thus  $g$  is a  $1 \times 1$  matrix given by*

$$\left( \frac{\partial}{\partial u} \cos u \right) \cdot \left( \frac{\partial}{\partial u} \cos u \right) = \sin^2 u + \cos^2 u = 1$$

*For any  $p \in \mathbb{S}^1$ , the tangent space  $T_p\mathbb{S}^1$  is the line tangent to the circle and orthogonal to the vector  $p$ . We thus have  $T_p\mathbb{S}^1$  is isomorphic to  $\mathbb{R}$  ( $T_p\mathbb{S}^1 \cong \mathbb{R}$ ). If we set  $p = (\cos u, \sin u)$ , we can represent a vector  $v \in T_p\mathbb{S}^1$  via  $v_0 \in \mathbb{R}$  and the relation  $v = v_0(-\sin u, \cos u)$ . This Riemannian metric makes sense as an inner product on this line now comes down to multiplying two real numbers representing a tangent vector on the line orthogonal to  $p$ .*

**Example 2.1.19** (The sphere  $\mathbb{S}^2$ ). *As we have seen previously, the sphere can be parameterized by spherical coordinates via the map  $(u, v) \mapsto (\cos u \sin v, \sin u \sin v, \cos v)$ . Again, we inherit the dot product from  $\mathbb{R}^3$  and thus, by setting  $x(u, v) = (\cos u \sin v, \sin u \sin v, \cos v) \in \mathbb{S}^2$ , we find that a Riemannian metric on  $\mathbb{S}^2$  at  $x$  is given by the matrix*

$$(g_{ij}) = \begin{pmatrix} \frac{\partial x}{\partial u} \cdot \frac{\partial x}{\partial u} & \frac{\partial x}{\partial u} \cdot \frac{\partial x}{\partial v} \\ \frac{\partial x}{\partial v} \cdot \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \cdot \frac{\partial x}{\partial v} \end{pmatrix} = \begin{pmatrix} \sin^2 v & 0 \\ 0 & 1 \end{pmatrix} \quad (2.3)$$

### 2.1.4. The Levi-Civita connection

We continue to let  $\mathcal{M}$  be a  $d$ -dimensional smooth manifold. Then for any two distinct points  $p, q \in \mathcal{M}$ , the tangent spaces  $T_p\mathcal{M}$  and  $T_q\mathcal{M}$  are  $d$ -dimensional vector spaces and thus isomorphic to  $\mathbb{R}^d$ . Hence they are isomorphic to each other. This isomorphism however, is generally far from trivial, but we can induce a relation between the two tangent planes using curves between  $p$  and  $q$ .

**Definition 2.1.20** (Connection). *Let  $\mathcal{M}$  be a smooth manifold and let  $\Gamma(T\mathcal{M})$  denote the set of all smooth vector fields on  $\mathcal{M}$ . A connection on  $T\mathcal{M}$  is a map*

$$\nabla : \Gamma(T\mathcal{M}) \times \Gamma(T\mathcal{M}) \rightarrow \Gamma(T\mathcal{M}) \quad (2.4)$$

denoted by  $(X, Y) \mapsto \nabla_X Y$  that satisfies the following properties

(i)  $\nabla_{fX+gY}Z = f\nabla_X Z + g\nabla_Y Z$  for all functions  $f, g : \mathcal{M} \rightarrow \mathbb{R}$  of class  $\mathcal{C}^\infty$ .

(ii)  $\nabla_X(aY + bZ) = a\nabla_X Y + b\nabla_X Z$  for all  $a, b \in \mathbb{R}$ .

(iii)  $\nabla_X(fY) = f\nabla_X Y + X(f)Y$  for all functions  $f : \mathcal{M} \rightarrow \mathbb{R}$  of class  $\mathcal{C}^\infty$ .

It is important to notice that a connection is a function that takes two vector fields as input and returns another one. The purpose of this operation is that connections give an insight in how two vector fields interact with each other by giving us an idea of how  $X$  moves over into  $Y$ . Now note that over a coordinate patch, the connection is uniquely defined by its values for  $X = \partial_j$  and  $Y = \partial_k$  as they form a basis for the tangent spaces. Since, by definition,  $\nabla_X Y$  is a vector field in  $\mathcal{M}$ , there must locally be smooth functions  $\Gamma_{jk}^i : \mathcal{M} \rightarrow \mathbb{R}$  such that

$$\nabla_{\partial_j} \partial_k = \Gamma_{jk}^i \partial_i \quad (2.5)$$

**Definition 2.1.21** (Christoffel symbols). *The functions  $\Gamma_{jk}^i$  of equation (2.5) are called the Christoffel symbols of the connection  $\nabla$ .*

**Proposition 2.1.22.** *Let  $(\mathcal{M}, g)$  be a Riemannian manifold, then there exists a unique affine connection  $\nabla$  that is compatible with the metric  $g$ , i.e. it satisfies the following conditions*

(i)  $\nabla g \equiv 0$

(ii) For all  $X, Y \in \Gamma(T\mathcal{M})$ ,  $[X, Y] = \nabla_X Y - \nabla_Y X$ .

*Proof.* See lemma 6.8 in Spivak, 1999, [26]. □

The notion  $\nabla g \equiv 0$  should be interpreted as follows. When writing  $g = \langle \cdot, \cdot \rangle$ , it should be clear that  $\langle X, Y \rangle$  defines a smooth map  $\mathcal{M} \rightarrow \mathbb{R}$  via  $p \mapsto \langle X_p, Y_p \rangle$ . We thus define  $\nabla_X \langle Y, Z \rangle = X(\langle Y, Z \rangle)$ , which defines a  $\mathcal{C}^\infty(\mathcal{M})$ -linear transformation  $X \mapsto \nabla_X g$ . It can be shown (See e.g. Spivak, 1999 [26] or Lovett, 2010 [14]) that  $\nabla g$  is identically 0 if and only if

$$\nabla_X \langle Y, Z \rangle = \langle \nabla_X Y, Z \rangle + \langle Y, \nabla_X Z \rangle. \quad (2.6)$$

This demonstrates that a connection that is compatible with the metric  $g$  satisfies the product rule with respect to the metric.

**Definition 2.1.23** (Levi-Civita connection). *The connection defined in proposition 2.1.22 is called the Levi-Civita connection.*

**Theorem 2.1.24.** *Let  $(\mathcal{M}, g)$  be a smooth Riemannian manifold, then over a coordinate patch of  $\mathcal{M}$  with coordinates  $(x^1, \dots, x^d)$ , the Christoffel symbols of the Levi-Civita connection are given by*

$$\Gamma_{jk}^i = \frac{1}{2} g^{i\ell} (\partial_j g_{k\ell} + \partial_k g_{j\ell} - \partial_\ell g_{jk}) \quad (2.7)$$

*Proof.* See proposition 5.2.13 of Lovett, 2010 [14] □

**Example 2.1.25** (The sphere  $\mathbb{S}^2$ ). We parameterize the sphere by  $(u, v) \mapsto (\cos u \sin v, \sin u \sin v, \cos v)$ . Note that equation (2.3) implies that

$$(g^{ij}) = \begin{pmatrix} \frac{1}{\sin^2 v} & 0 \\ 0 & 1 \end{pmatrix} \quad (2.8)$$

We now substitute equations (2.3) and (2.8) into (2.7), where  $\partial_1 = \frac{\partial}{\partial u}$  and  $\partial_2 = \frac{\partial}{\partial v}$ . Standard calculations then show that the Christoffel symbols for the sphere are given by

$$\begin{aligned} \Gamma_{11}^1 &= 0 & \Gamma_{11}^2 &= -\sin v \cos v \\ \Gamma_{12}^1 &= \frac{1}{\tan v} & \Gamma_{12}^2 &= 0 \\ \Gamma_{21}^1 &= \frac{1}{\tan v} & \Gamma_{21}^2 &= 0 \\ \Gamma_{22}^1 &= 0 & \Gamma_{22}^2 &= 0 \end{aligned} \quad (2.9)$$

### 2.1.5. Vector fields along curves, geodesics and the exponential map

We now know how to define a vector field on a manifold, but we can also define a vector field along a curve.

**Definition 2.1.26** (Vector field along a curve). Let  $\gamma : I \rightarrow \mathcal{M}$  be a smooth curve in  $\mathcal{M}$ , where  $I \subseteq \mathbb{R}$  is an interval. We call  $V$  a vector field along  $\gamma$  if for each  $t \in I$ ,  $V(t)$  is a tangent vector in  $T_{\gamma(t)}\mathcal{M}$  and if  $V$  defines a smooth map  $I \rightarrow T\mathcal{M}$ .

It is important to notice that a vector field along  $\gamma$  need not be a restriction of a vector field  $X$  on  $\mathcal{M}$  to  $\gamma(I)$ . In order to clarify this, consider a curve that intersects with itself ( $\gamma(t_0) = \gamma(t_1)$  for some pair  $t_0 \neq t_1$  in  $I$ ). We can very well define a vector field along  $\gamma$  with  $V(t_0) \neq V(t_1)$ , but then there cannot exist a vector field  $X$  on  $\mathcal{M}$  with  $V(t) = X(\gamma(t))$  for all  $t \in I$ . If a vector field along a curve is a restriction of a vector field  $X$  to  $\gamma(I)$ , however, we say that  $V$  is induced from  $X$  or that  $V$  extends to  $X$ . We will denote the set of all smooth vector fields on  $\mathcal{M}$  along a curve  $\gamma$  by  $\mathcal{X}_\gamma(\mathcal{M})$ .

**Example 2.1.27.** The most elementary example of a vector field along a curve is its derivative. That is, let  $\gamma : I \rightarrow \mathcal{M}$  be a smooth curve, then at any time  $t$  we have  $\gamma'(t) \in T_{\gamma(t)}\mathcal{M}$ , thus the map  $t \mapsto \gamma'(t)$  defines a vector field along  $\gamma$ , since smoothness follows from smoothness of  $\gamma$ .

**Proposition 2.1.28.** Let  $\nabla$  denote the Levi-Civita connection. There exists a unique operator  $D_t : \mathcal{X}_\gamma(\mathcal{M}) \rightarrow \mathcal{X}_\gamma(\mathcal{M})$  such that

(i)  $D_t(V + W) = D_t V + D_t W$  for all  $V, W \in \mathcal{X}_\gamma(\mathcal{M})$ .

(ii)  $D_t(fV) = \frac{df}{dt}V + fD_t V$  for all  $V \in \mathcal{X}_\gamma(\mathcal{M})$  and  $f \in \mathcal{C}^\infty(I)$ .

(iii) If  $V$  extends to a vector field  $X \in \Gamma(T\mathcal{M})$ , then  $D_t V = \nabla_{\gamma'(t)} X$ .

*Proof.* Suppose the operator  $D_t$  exists and let  $x = (x^1, \dots, x^d)$  be local coordinates over a chart  $U$  of  $\mathcal{M}$ . Let  $V \in \mathcal{X}_\gamma(\mathcal{M})$  be a smooth vector field along  $\gamma$  and write  $V = v^i \partial_i$  on  $U$ , where  $v^i, i = 1, \dots, d$  are smooth functions on  $I$ . Conditions (i) and (ii) now tell us that we must have that

$$D_t V = \dot{v}^i \partial_i + v^i D_t \partial_i$$

where we use the notation  $\dot{v}^i$  for the  $i$ 'th component of  $v'$ . Now write  $\gamma = (\gamma^1, \dots, \gamma^d)$ . Then over  $U$ , we have  $\gamma' = \dot{\gamma}^i \partial_i$ , so condition (iii) and the properties of a connection state that

$$D_t \partial_i = \nabla_{\gamma'(t)} \partial_i = \dot{\gamma}^j \nabla_{\partial_j} \partial_i = \dot{\gamma}^j \Gamma_{ij}^k \partial_k$$

By rearranging the indices, we thus deduce that

$$D_t V = \left( \dot{v}^i + \Gamma_{jk}^i \dot{\gamma}^j v^k \right) \partial_i \quad (2.10)$$

which shows that if  $D_t$  does exist, it has to be unique as well. To proof existence, we define the operator  $D_t^\alpha$  as in equation (2.10) for each chart  $U_\alpha$ . From uniqueness of  $D_t^\alpha$  on  $U_\alpha$ , it follows that  $D_t^\alpha = D_t^\beta$  on  $U_\alpha \cap U_\beta$  for all overlapping charts. Since  $\alpha$  ranges over all coordinate charts in the atlas, the collection  $\{D_t^\alpha\}_\alpha$  thus extends to one unique operator on  $\mathcal{M}$ .  $\square$

The operator  $D_t$  is called the *covariant derivative along  $\gamma$* . We can observe that  $D_t V$  adjusts the derivative of the vector field  $V$  by correcting for the shape of the curve on which  $V$  is defined. A natural interpretation of the covariant derivative therefore is that it gives us an idea of how a vector field changes along a curve on  $\mathcal{M}$ , and from the Christoffel symbols belonging to the Levi-Civita connection that appear in equation (2.10), it seems that it also gives us an idea of how the Riemannian metric behaves along a curve. This becomes particularly clear when noticing that a combination of part (iii) of proposition 2.1.28 and equation (2.6) leads to

$$\frac{d}{dt}g(V, W) = g(D_t V, W) + g(V, D_t W). \quad (2.11)$$

**Definition 2.1.29** (Geodesic). *A curve  $\gamma : I \rightarrow \mathcal{M}$  is called a geodesic if it has acceleration 0, that is  $D_t \gamma'(t) = 0$  for all  $t \in I$ .*

Note that (2.11) now implies the physical interpretation that a geodesic also has constant speed, since the definition implies that  $\frac{d}{dt}g(\gamma'(t), \gamma'(t)) = 0$ . This result demonstrates that when a curve is parameterized by arc length, i.e. the length of the curve is the same as the parameterization  $t = \int_0^t \|\gamma'(s)\| ds$ , we must have that the speed  $g(\gamma'(t), \gamma'(t))$  is identically 1.

Geodesics are a very important topic when talking about Riemannian Manifolds, as these are the curves that do not have any change in velocity and thus follow the same direction. One can also show (see e.g. Lovett, 2010 [14] or Spivak, 1970 [25]) that the shortest curve between two points on a manifold will always be a geodesic, which intuitively makes sense as geodesics move can be interpreted to move in a straight line and gives us an alternative for the triangle inequality on manifolds by stating that a straight line between points is shorter than any other curve between the points. Before we give some examples of geodesics on some well-known manifolds, we have to derive a result that helps us to identify a curve as geodesic.

If we substitute  $\gamma'$  into equation (2.10) and apply the definition of geodesics 2.1.29, we can derive the *Geodesic Equations*, stating that a curve  $\gamma$  on  $\mathcal{M}$  is a geodesic on a chart  $U$  with coordinate system  $x = (x^1, \dots, x^d)$  if and only if the following equations are satisfied.

$$\frac{d^2 \gamma^i}{dt^2} + \Gamma_{jk}^i(\gamma(t)) \frac{d\gamma^j}{dt} \frac{d\gamma^k}{dt} = 0 \quad \text{for all } i \quad (2.12)$$

Here  $\gamma(t) = (\gamma^1(t), \dots, \gamma^d(t))$ . It is important to notice that these equations form a system of  $d$  nonlinear differential equations and therefore it is usually difficult to find a closed form for the geodesics on a manifold. The following theorem now follows from standard results from ordinary differential equations.

**Theorem 2.1.30.** *Let  $p \in \mathcal{M}$ . For any  $v \in T_p \mathcal{M}$  and any  $t_0 \in \mathbb{R}$ , there exists an open interval  $I$  with  $t_0 \in I$  and a unique geodesic  $\gamma : I \rightarrow \mathcal{M}$  with  $\gamma(t_0) = p$  and  $\gamma'(t_0) = v$ .*

*Proof.* See theorem 5.3.12 in Lovett, 2010 [14].  $\square$

Theorem 2.1.30 shows us that we can start at any point in  $\mathcal{M}$  and walk in the direction of  $v \in T_p \mathcal{M}$  with constant velocity and thus motivates a definition of the exponential map.

**Definition 2.1.31** (The exponential map). *Let  $p \in \mathcal{M}$ . The exponential map  $\text{Exp}_p : T_p \mathcal{M} \rightarrow \mathcal{M}$  is defined at any  $v \in T_p \mathcal{M}$  as the unique geodesic  $\gamma^v$  from theorem 2.1.30 evaluated at  $t = 1$ , i.e.  $\text{Exp}_p v = \gamma^v(1)$ .*

**Definition 2.1.32** (Logarithm and geodesic distance). *The inverse map  $\text{Log}_p : \mathcal{M} \rightarrow T_p \mathcal{M}$  assigns to a point  $q \in \mathcal{M}$  the vector  $v \in T_p \mathcal{M}$  such that the geodesic  $\gamma^v$  starting at  $p$  with initial velocity  $v$  satisfies  $\gamma^v(1) = q$ . The logarithmic map induces the geodesic distance on  $\mathcal{M}$  given by  $d(p, q) = \|\text{Log}_p q\|$  that generalizes the idea that the distance between two points is given by the length of the shortest curve connecting them. The norm is understood to be a Euclidean norm, which can be taken as  $T_p \mathcal{M}$  is isomorphic to a Euclidean space.*

**Example 2.1.33** (The exponential map on the circle). A geodesic on the circle at a given time  $t$  is given by the vector  $S(t) = (\cos \psi(t), \sin \psi(t))$ . Since we try to find  $\text{Exp}_p v$ , we must have that  $S(0) = p$  and  $S'(0) = v$ . Let us parameterize the point  $p$  by  $p = (\cos \theta, \sin \theta)$  for some  $\theta \in [0, 2\pi]$ . Then we must have that  $\psi(0) = \theta$ , which implies that  $S'(0) = \psi'(0) (-\sin \theta, \cos \theta)$ . Since  $v \in T_p \mathbb{S}^1$  example 2.1.9 tells us that  $v$  must represent an orthogonal vector to  $p$  in  $\mathbb{R}^2$  and thus we immediately have that it is of the form  $v = v_0 (-\sin \theta, \cos \theta)$ , where  $v_0 \in \mathbb{R}$  represents its magnitude. Hence  $\psi'(0) = v_0$  (here a positive sign means a counterclockwise movement). We can thus satisfy the requirements by taking  $\psi(t) = \theta + tv_0$  and thus  $\text{Exp}_p v = (\cos(\theta + v_0 t), \sin(\theta + v_0 t))$ . This means that we move on the circle either clockwise or counterclockwise (depending on the sign of  $v_0$ ) where the distance we move is determined by the magnitude of  $v_0$ .

**Example 2.1.34** (Geodesics on the sphere  $\mathbb{S}^2$ ). We again consider the unit sphere under the parameterization  $(u, v) \mapsto (\cos u \sin v, \sin u \sin v, \cos v)$ . A curve on the sphere can now be parameterized by  $\gamma(t) = (u(t), v(t))$ . Using the Christoffel symbols found in equation (2.9), we find that the geodesic equations on the sphere read

$$\begin{aligned} \frac{d^2 u}{dt^2} + \frac{2}{\tan v} \frac{du}{dt} \frac{dv}{dt} &= 0 \\ \frac{d^2 v}{dt^2} - \sin v \cos v \left( \frac{du}{dt} \right)^2 &= 0 \end{aligned} \quad (2.13)$$

Since a geodesic is written in the form  $\gamma(t) = (\cos u(t) \sin v(t), \sin u(t) \sin v(t), \cos v(t))$  we can find that

$$\gamma'(t) = \left( -\sin u \sin v \frac{du}{dt} + \cos u \cos v \frac{dv}{dt}, \cos u \sin v \frac{du}{dt} + \sin u \cos v \frac{dv}{dt}, -\sin v \frac{dv}{dt} \right).$$

Taking a second derivative and applying the equations found in (2.13), we find after some simplification that

$$\gamma''(t) = - \left( \sin^2 v \left( \frac{du}{dt} \right)^2 + \left( \frac{dv}{dt} \right)^2 \right) \gamma(t)$$

Now notice that  $\left( \sin^2 v \left( \frac{du}{dt} \right)^2 + \left( \frac{dv}{dt} \right)^2 \right) = g(\gamma'(t), \gamma'(t))$  is the speed of the curve squared, which is constant as  $\gamma$  is a geodesic. We therefore find that  $\gamma$  satisfies

$$\gamma''(t) + C\gamma(t) = 0$$

where  $C$  is a positive constant such that  $\sqrt{C}$  equals the velocity of  $\gamma$ . Thus the geodesics on the sphere can be found by

$$\gamma(t) = a \cos(\sqrt{C}t) + b \sin(\sqrt{C}t) \quad (2.14)$$

where  $a, b$  are vectors in  $\mathbb{R}^3$ . One quickly calculates that  $\gamma(0) = a$  and  $\gamma'(0) = \sqrt{C}b$ . Since  $\gamma$  must lie on the sphere at all times, we also require that  $\|a\| = \|b\| = 1$  and  $a \cdot b = 0$ . The geodesics on the sphere are thus indeed found by the circles around it. One can choose the initial point  $a$  and the initial velocity  $\sqrt{C}b$  freely

**Corollary 2.1.35** (Exponential map on the sphere). The exponential map on the sphere for  $p \in \mathbb{S}^2$  is given by

$$\text{Exp}_p v = \begin{cases} p & \text{if } v = 0 \\ \cos(\|v\|)p + \sin(\|v\|) \frac{v}{\|v\|} & \text{if } \|v\| > 0 \end{cases} \quad (2.15)$$

*Proof.* The form of the geodesics is given in equation (2.14). Since the exponential map should evaluate the unique geodesic  $\gamma$  from theorem 2.1.30 at  $t = 1$ . This geodesic satisfies  $\gamma(0) = p$  and  $\gamma'(0) = v$  and we thus find that  $a = \gamma(0) = p$  and  $b\sqrt{C} = \gamma'(0) = v$ . Equation (2.15) now follows as we must have that  $\|b\| = 1$ .  $\square$

### 2.1.6. Gradients, divergence and the Laplace-Beltrami operator

We finish this section on Riemannian geometry by extending some well-known differential operators to operators on a Riemannian manifold  $(\mathcal{M}, g)$ .

Recall that in  $\mathbb{R}^d$ , the gradient vector of a function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  defines the vector field  $\nabla f = (\partial_1 f, \dots, \partial_d f)$  and is the unique vector that solves  $\nabla f(x) \cdot v = D_v f(x)$  for any vector  $v \in \mathbb{R}^d$ . Here  $D_v f(x)$  denotes the directional derivative of  $f$  along  $v$  evaluated in  $x$ . On a Riemannian manifold  $\mathcal{M}$  we use a similar construction.

**Definition 2.1.36** (Gradient). *The gradient  $\text{grad}f$  of a differentiable function  $f : \mathcal{M} \rightarrow \mathbb{R}$  is the unique vector field that satisfies*

$$g(\text{grad}f, X) = \text{d}f(X) = Xf \quad (2.16)$$

for all vector fields  $X : \mathcal{M} \rightarrow T\mathcal{M}$ .

Now let  $(x^1, \dots, x^d)$  be a local coordinate system and set  $X = X^i \partial_i$ . Then  $\text{grad}f$  must be of the form  $\text{grad}f = a^i \partial_i$ . Substituting this into equation (2.16) yields

$$X^j \partial_j f = g(\text{grad}f, X) = g_{ij} a^i X^j$$

From which it follows that for all  $j$ , we must have  $\partial_j f = g_{ij} a^i$ , which leads us to

$$a^k = \delta_{jk} a^j = g^{ik} g_{ji} a^j = g^{ik} \partial_i f$$

where  $\delta$  is the Kronecker delta function. Note that the second identity here follows from  $(g^{ij})(g_{ij}) = I$  and the symmetry of the matrix  $g$ . This leads us to a local characterization of the gradient on  $(\mathcal{M}, g)$  given by

$$\text{grad}f = g^{ij} \partial_i f \partial_j$$

The next operator we consider in this section is the divergence operator. The divergence  $\text{div}X$  of a vector field  $X : \mathcal{M} \rightarrow T\mathcal{M}$  on  $\mathcal{M}$  turns the vector field into a scalar field. In  $\mathbb{R}^d$  this happens via  $\text{div}X = \partial_j X^j$ . Just like the gradient however, this works a bit differently when working on differentiable manifolds.

**Definition 2.1.37** (Divergence of a vector field). *Let  $(\mathcal{M}, g)$  be a smooth manifold. The divergence of a vector field  $X : \mathcal{M} \rightarrow T\mathcal{M}$  is defined as  $\text{div}X = \frac{1}{\sqrt{|g|}} \partial_j (\sqrt{|g|} X^j)$  in local coordinates where  $|g|$  denotes the determinant of the matrix  $(g_{ij})$ .*

**Theorem 2.1.38.** *The divergence operator can also be written as*

$$\partial_j X^j + \Gamma_{ij}^i X^j \quad (2.17)$$

In order to prove this theorem, we first introduce a result from linear algebra: Jacobi's formula:

**Lemma 2.1.39** (Jacobi's formula for invertible matrices). *Let  $A : \mathbb{R} \rightarrow \mathbb{R}^{d \times d}$  be a differentiable map such that  $A(t)$  is invertible for all  $t \in \mathbb{R}$ . Then the determinant defines a differentiable map  $\mathbb{R} \rightarrow \mathbb{R}$  such that*

$$\frac{\text{d}}{\text{d}t} |A(t)| = \text{Trace} \left( |A(t)| A(t)^{-1} \frac{\text{d}}{\text{d}t} A(t) \right) \quad (2.18)$$

*Proof.* Without loss of generality, we will show that the formula holds in case  $t = 0$ . We first show the result in a very specific case:  $A(t) = tA + I$ , where  $A$  is a constant matrix. In this case

$$\frac{\text{d}}{\text{d}t} |A(t)| = \lim_{t \rightarrow 0} \frac{|I + tA| - |I|}{h}$$

Now notice that  $|I + tA|$  is a polynomial in  $t$  of order  $d$  that has constant term 1. The linear terms of the polynomial only appear when the multiplication happens between terms on the diagonal (as the  $+1$  appears there) so the linear term must be  $\text{Trace}(A)$ . Therefore

$$\left. \frac{\text{d}}{\text{d}t} |A(t)| \right|_{t=0} = \text{Trace}(A) = \text{Trace} \left( |A(0)| A(0)^{-1} \frac{\text{d}}{\text{d}t} A(t) \right) \Big|_{t=0}$$

Now let us go back to the more general case where  $A$  is an invertible matrix. We can write

$$A(t) = A(0) + tA'(0) + \mathcal{O}(t^2)$$

And thus we must have that

$$\begin{aligned} \left. \frac{d}{dt} |A(t)| \right|_{t=0} &= \lim_{t \rightarrow 0} \frac{|A(t)| - |A(0)|}{t} \\ &= \lim_{t \rightarrow 0} \frac{|A(0) + tA'(0)| - |A(0)|}{t} + \mathcal{O}(t) \\ &= \lim_{t \rightarrow 0} \frac{|A(0)| |I + tA(0)^{-1}A'(0)| - |A(0)|}{t} \\ &= |A(0)| \text{Trace} \left( A(0)^{-1}A'(0) \right) \end{aligned}$$

which is precisely what we needed to show.  $\square$

*Proof of theorem 2.1.38.* By applying both the product and the chain rule, one obtains that

$$\text{div} X = \frac{1}{\sqrt{|g|}} \partial_j \left( \sqrt{|g|} X^j \right) = \partial_j X^j + \frac{1}{2|g|} X^j \partial_j |g|$$

We can now apply lemma 2.1.39 to find that

$$\text{div} X = \partial_j X^j + \frac{1}{2} X^j \text{Trace} (g^{-1} \partial_j g) = \partial_j X^j + \frac{1}{2} X^j g^{ik} \partial_j g_{ki}$$

Now all that is left is to notice that by symmetry of the matrix  $(g_{ij})$ , we have

$$\Gamma_{ij}^i = \frac{1}{2} g^{ik} (\partial_i g_{jk} + \partial_j g_{ki} - \partial_k g_{ij}) = \frac{1}{2} g^{ik} \partial_j g_{ki}$$

$\square$

Now that we have equivalents for the divergence and the gradient operators on manifolds, we can define an equivalent for the Laplace operator as well. Recall that in  $\mathbb{R}^d$  the Laplace operator is given by  $\Delta = \text{div} \nabla = \sum_{i=1}^d \frac{\partial}{\partial (x^i)^2}$ .

**Definition 2.1.40** (Laplace-Beltrami operator). *Let  $(\mathcal{M}, g)$  be a Riemannian manifold. The Laplace-Beltrami operator on a smooth function  $f : \mathcal{M} \rightarrow \mathbb{R}$  is defined as  $\Delta_{\mathcal{M}} f = \text{div} \text{grad} f$*

Using the definitions of divergence and the gradient, it is not hard to see that in local coordinates, the Laplace-Beltrami operator can be expressed as

$$\Delta_{\mathcal{M}} f = \frac{1}{\sqrt{|g|}} \partial_j \left( \sqrt{|g|} g^{ij} \partial_i f \right) \quad (2.19)$$

A different expression for the Laplace-Beltrami operator can be found using theorem 2.1.38 as

$$\text{div} \text{grad} f = \text{div} (g^{ij} \partial_i f \partial_j) = \partial_j g^{ij} \partial_i f + \Gamma_{ij}^i g^{ij} \partial_i f = g^{ij} \partial_i \partial_j f + (\partial_j g^{ij} + \Gamma_{ij}^i g^{ij}) \partial_i f \quad (2.20)$$

Now note that we have the identity

$$0 = \partial_j I = \partial_j (g g^{-1}) = g \partial_j g^{-1} + (\partial_j g) g^{-1}$$

which leads to  $\partial_j g^{-1} = -g^{-1} (\partial_j g) g^{-1}$ . Plugging this in into equation (2.20) and rearranging the terms results in

$$\Delta_{\mathcal{M}} f = g^{ij} \partial_i \partial_j f - g^{jk} \Gamma_{jk}^i \partial_i f \quad (2.21)$$

**Example 2.1.41** (Euclidean spaces). *Note that on Euclidean spaces with the standard dot product, we have that  $(g_{ij}) = I$  and we thus have that  $\text{grad} f = \sum_i \partial_i f \partial_i$  and  $\text{div} X = \partial_j X^j$ , which leads to  $\Delta f = \sum_i \partial_i^2 f$ . The operators on manifolds thus reduce to the classical formulas for these operators on Euclidean spaces.*

**Example 2.1.42** (The circle  $\mathbb{S}^1$ ). *On the circle parameterized by  $u \mapsto (\cos u, \sin u)$  we have*

$$(\Delta_{\mathbb{S}^1} f)(u) = \frac{\partial^2 f}{\partial u^2}$$

**Example 2.1.43** (The sphere  $\mathbb{S}^2$ ). *Recall equation (2.3). Filling these values in into equations (2.5) and (2.21), results in*

$$(\Delta_{\mathbb{S}^2} f)(u, v) = \frac{1}{\sin^2 v} \frac{\partial^2 f}{\partial u^2} + \frac{\partial^2 f}{\partial v^2} + \frac{1}{\tan v} \frac{\partial f}{\partial v} \quad (2.22)$$

## 2.2. Stochastic processes: Itô and Stratonovich calculus

In this section, we introduce diffusion processes and stochastic differential equations. Many of the notions introduced in this section are based on the lecture notes by Spieksma and van Zanten, 2017 [24], the notes by Eberle, 2016 [7] and the books on stochastic calculus by Cohen and Elliot, 2015 [5] and Øksendal, 2003 [17].

### 2.2.1. Stochastic processes

Throughout, it will be assumed that there is an underlying filtered probability space  $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$  that satisfies the following *usual conditions* :

- The probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  is a complete probability space.
- $(\mathcal{F}_t)_{t \geq 0}$  is a filtration of sub- $\sigma$ -fields of  $\mathcal{F}$  is right-continuous. That is,  $\mathcal{F}_t$  is a  $\sigma$ -field for all  $t \geq 0$ ,  $\emptyset \in \mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}$  for all  $s < t$  and  $\mathcal{F}_t = \mathcal{F}_{t+} := \bigcap_{s > t} \mathcal{F}_s$  for all  $t$ .
- $\mathcal{F}_0$  contains all  $\mathbb{P}$ -negligible events of  $\mathcal{F}_\infty := \sigma(\mathcal{F}_t : t \geq 0)$

A random element of a measurable space  $(\mathcal{X}, \mathcal{B})$  is defined as a measurable map  $X : (\Omega, \mathcal{F}) \rightarrow (\mathcal{X}, \mathcal{B})$ . A stochastic process is a family  $X = (X_t)_{t \geq 0}$  of random elements of  $(\mathcal{X}, \mathcal{B})$ . We call a stochastic process continuous if  $\mathbb{P}$ -almost surely, each of its *sample paths*  $\omega \mapsto X(\omega)$  is a continuous function of  $t$ . The process is adapted to the filtration if each map  $X_t : (\Omega, \mathcal{F}_t) \rightarrow (\mathcal{X}, \mathcal{B})$  is measurable.

Recall that a random variable  $X : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (\mathbb{R}^d, \mathcal{B})$  defines a probability measure  $\mathbb{P}_X$  on  $\mathbb{R}^d$  by setting  $\mathbb{P}_X(A) = \mathbb{P}(X^{-1}(A))$ . The same can be done for a stochastic process.

**Definition 2.2.1** (Law of a stochastic process). *Let  $\mathcal{X}^{[0, T]}$  denote the set of all functions  $[0, T] \rightarrow \mathcal{X}$  for a given  $T > 0$ . When  $\mathcal{X}^{[0, T]}$  is equipped with a suitable  $\sigma$ -field, a process  $X = (X_t : t \in [0, T])$  defines a measurable map  $\Omega \rightarrow \mathcal{X}^{[0, T]}$  given by  $\omega \mapsto X(\omega)$  and we then induce a probability measure on  $\mathcal{X}^{[0, T]}$  which can, with slight abuse of notation be denoted as  $\mathbb{P}_T(A) = \mathbb{P}(X^{-1}(A))$  for  $A \subseteq \mathcal{X}^{[0, T]}$ . This measure will be referred to as the law of the process  $X$ .*

In practice, the  $\sigma$ -field on  $\mathcal{X}^{[0, T]}$  the  $\sigma$ -field generated by the so-called  $\sigma$ -cylinders and the existence of a law traces back to the existence of finite dimensional distributions, i.e. the distributions of the  $\mathcal{X}^k$ -valued random vectors  $(X_{t_1}, \dots, X_{t_k})$  for any set of times  $0 \leq t_1, \dots, t_k \leq T$ . See Spieksma and van Zanten, 2017 [24] for further reading.

**Definition 2.2.2** (Independence of stochastic processes). *We say that two stochastic processes are independent if they define independent random elements of the space  $\mathcal{X}^{[0, T]}$ . It follows (see e.g. Spieksma and van Zanten, 2017 [24]) that Independence of stochastic processes is equivalent to independence in the finite dimensional distributions.*

**Definition 2.2.3** (Brownian motion in  $\mathbb{R}$ ). *A special case of a stochastic process is the Brownian motion, also frequently referred to as the Wiener process. A Brownian motion  $W = (W_t)_t$  with initial distribution  $\nu$  is a stochastic process adapted to a filtration  $(\mathcal{F}_t)_t$  that satisfies the following properties*

- (i) *The random variable  $W_0$  is  $\mathcal{F}_0$ -measurable and has probability distribution  $\nu$ .*
- (ii)  *$W$  is almost surely continuous, i.e. the function  $t \mapsto W_t(\omega)$  is continuous for  $\mathbb{P}$ -almost all  $\omega \in \Omega$ .*
- (iii) *For all  $s, t \geq 0$  with  $s < t$ , the increment  $W_t - W_s$  is independent of  $\mathcal{F}_s$  and follows a  $\mathcal{N}(0, t - s)$ -distribution.*

When  $\nu$  is such that  $W_0 = 0$   $\nu$ -almost surely, we refer to  $W$  as a standard Brownian motion.

Notice that from (ii) and (iii) it follows that, for a Brownian motion  $W$  starting at  $x \in \mathbb{R}$ ,  $W_t = W_t - W_0 + W_0$  and thus

$$\mathbb{P}(W_t \in A) = \int_A \frac{1}{\sqrt{2\pi t}} e^{-\frac{(y-x)^2}{2t}} dy, \quad A \in \mathcal{B}(\mathbb{R}). \quad (2.23)$$

It is well-known that such a process as defined in definition 2.2.3 exists and its sample paths are almost surely nowhere differentiable (An elegant proof can be seen in Kallenberg, 2006 [10]). An example of a simulation of a path of Brownian Motion starting at 0 is given in figure 2.2.

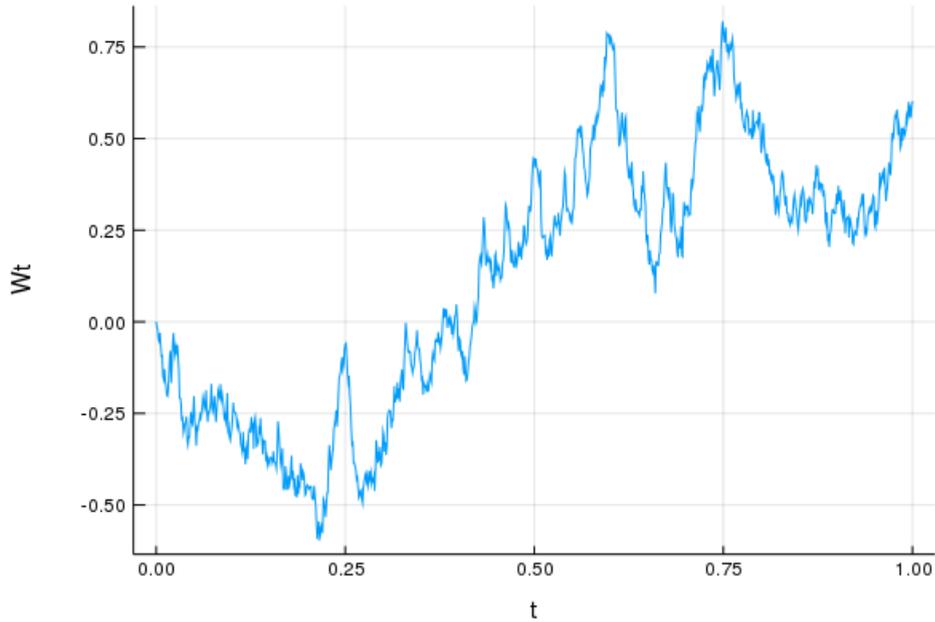


Figure 2.2: A simulated path of a standard Brownian Motion in  $\mathbb{R}$  starting at 0.

An important property of Brownian motion is the martingale property. In appendix A, we give some background theory on martingales and demonstrate that Brownian motion is indeed a martingale process.

**Definition 2.2.4** (Brownian motion in  $\mathbb{R}^d$ ). *A  $d$ -dimensional Brownian motion is an  $\mathbb{R}^d$ -valued stochastic process  $W = (W_t^1, \dots, W_t^d)_t$  where  $W^1, \dots, W^d$  are independent  $\mathbb{R}$ -valued Brownian motions. Examples of simulated paths in  $\mathbb{R}^2$  and  $\mathbb{R}^3$  can be found below.*

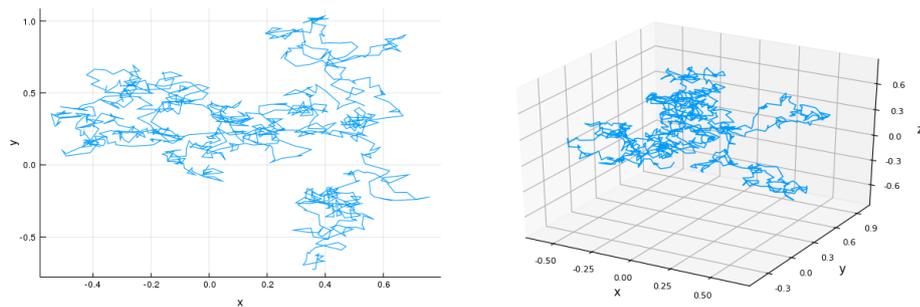


Figure 2.3: Simulations of realizations of  $\mathbb{R}^2$ -valued Brownian motion and  $\mathbb{R}^3$ -valued Brownian motion both starting at 0.

We also make use of quadratic variation and quadratic covariation of stochastic processes  $X = (X_t)_t$  and  $Y = (Y_t)_t$ .

**Definition 2.2.5** (Quadratic (co)variation). *The quadratic variation of  $X$  is the process  $[X]_t$  defined as*

$$[X]_t := \lim_{\|\Delta\| \rightarrow 0} \sum_{k=1}^N (X_{t_k} - X_{t_{k-1}})^2, \quad t > 0 \quad (2.24)$$

where  $\Delta$  ranges over partitions of  $[0, t]$  and  $\|\Delta\| := \max\{|t_i - t_{i-1}| : i = 1, \dots, N\}$  and the limit is taken as a limit in probability. If  $X = (X_t)_t$  and  $Y = (Y_t)_t$  are two stochastic processes, the quadratic covariation of  $X$  and  $Y$  is defined as

$$[X, Y]_t := \lim_{\|\Delta\| \rightarrow 0} \sum_{k=1}^N (X_{t_k} - X_{t_{k-1}}) (Y_{t_k} - Y_{t_{k-1}}), \quad t > 0. \quad (2.25)$$

In appendix A, we elaborate more on quadratic variation and introduce some results in detail.

**Proposition 2.2.6.** *For the quadratic covariation, the following identities can be found by rewriting the terms in the summation.*

$$[X, Y]_t = \frac{[X + Y]_t - [X - Y]_t}{4} = \frac{[X + Y]_t - [X]_t - [Y]_t}{2}$$

In appendix A, we also deduce that the quadratic covariance defines a bilinear map over  $X$  and  $Y$ .

*Proof.* Let  $\Delta$  range over all partitions  $\{0 = t_0 < \dots < t_N = t\}$  of  $[0, t]$ . For the first identity, notice that linearity of limits and summations implies that

$$\begin{aligned} \frac{1}{4} ([X + Y]_t - [X - Y]_t) &= \frac{1}{4} \lim_{\|\Delta\| \rightarrow 0} \sum_{k=1}^N \left\{ (X_{t_k} + Y_{t_k} - X_{t_{k-1}} - Y_{t_{k-1}})^2 - (X_{t_k} - Y_{t_k} - X_{t_{k-1}} + Y_{t_{k-1}})^2 \right\} \\ &= \frac{1}{4} \lim_{\|\Delta\| \rightarrow 0} \sum_{k=1}^N \left\{ (X_{t_k} - X_{t_{k-1}})^2 + (Y_{t_k} - Y_{t_{k-1}})^2 + 2(X_{t_k} - X_{t_{k-1}})(Y_{t_k} - Y_{t_{k-1}}) \right. \\ &\quad \left. - (X_{t_k} - X_{t_{k-1}})^2 - (Y_{t_{k-1}} - Y_{t_k})^2 - 2(X_{t_k} - X_{t_{k-1}})(Y_{t_{k-1}} - Y_{t_k}) \right\} \\ &= \lim_{\|\Delta\| \rightarrow 0} \sum_{k=1}^N (X_{t_k} - X_{t_{k-1}}) (Y_{t_k} - Y_{t_{k-1}}) = [X, Y]_t \end{aligned}$$

Similarly, we have that

$$\begin{aligned} \frac{1}{2} ([X + Y]_t - [X]_t - [Y]_t) &= \frac{1}{2} \lim_{\|\Delta\| \rightarrow 0} \sum_{k=1}^N \left\{ (X_{t_k} + Y_{t_k} - X_{t_{k-1}} - Y_{t_{k-1}})^2 - (X_{t_k} - X_{t_{k-1}})^2 - (Y_{t_{k-1}} - Y_{t_k})^2 \right\} \\ &= \frac{1}{2} \lim_{\|\Delta\| \rightarrow 0} \sum_{k=1}^N \left\{ (X_{t_k} - X_{t_{k-1}})^2 + (Y_{t_{k-1}} - Y_{t_k})^2 + 2(X_{t_k} - X_{t_{k-1}})(Y_{t_{k-1}} - Y_{t_k}) \right. \\ &\quad \left. - (X_{t_k} - X_{t_{k-1}})^2 - (Y_{t_{k-1}} - Y_{t_k})^2 \right\} \\ &= \lim_{\|\Delta\| \rightarrow 0} \sum_{k=1}^N (X_{t_k} - X_{t_{k-1}}) (Y_{t_k} - Y_{t_{k-1}}) = [X, Y]_t \end{aligned}$$

□

## 2.2.2. Stochastic integration and stochastic differential equations

As this thesis is largely about stochastic differential equations, we will now define stochastic integration in a proper manner in order to move to this theory. For a detailed construction of the Itô integral, we refer to Kallenberg, 2006 [10], in this thesis we will stick with the following definition.

**Definition 2.2.7** (Itô integral). *For a continuous adapted stochastic process  $(H_t)_t$  and a semimartingale  $(X_t)_t$  and a sequence  $\{\pi_n\}_n$  of partitions of  $[0, \infty)$  with  $\|\pi_n\| \rightarrow 0$  as  $n \rightarrow \infty$ , the Itô integral of  $H$  with respect to  $X$  is defined as the process*

$$(H \cdot X)_t = \int_0^t H_s dX_s = \lim_{n \rightarrow \infty} \sum_{s \in \pi_n} H_s (X_{s' \wedge t} - X_{s \wedge t}) \quad (2.26)$$

where  $s' := \min\{u \in \pi : u > s\}$  whenever this limit exists and is  $L_2$ -bounded. The limit in equation (2.26) understood to be a limit in  $L^2(\Omega)$ .

At first glance, it is not clear whether or not this integral exists. There is a lot of literature available that shows that this is indeed the case (see e.g. [24], [7] and [17]). Moreover, the limit does not depend on the choice of partitions  $(\pi_n)$  with  $\|\pi_n\| \rightarrow 0$ .

**Corollary 2.2.8** (Linearity of the Itô integral). *If  $H^1$  and  $H^2$  are both continuous adapted processes and  $\alpha$  and  $\beta$  constants, then so is  $\alpha H^1 + \beta H^2$  and  $((\alpha H^1 + \beta H^2) \cdot X) = \alpha(H^1 \cdot X) + \beta(H^2 \cdot X)$*

Now that we know how to write a stochastic integral, we can consider processes defined as solutions to

$$X_t = X_0 + \int_0^t b(s, X_s) ds + \int_0^t \sigma(s, X_s) dW_s$$

where  $b, \sigma \in \mathcal{C}([0, \infty) \times \mathbb{R})$  and  $W$  denotes a Brownian motion. Equations of this form are generally written briefly in the form of stochastic differential equations (SDEs)

$$dX_t = b(t, X_t) dt + \sigma(t, X_t) dW_t. \quad (2.27)$$

The functions  $b$  and  $\sigma$  in equation (2.27) are frequently called the drift and the diffusion coefficient respectively of the process  $X = (X_t)_t$ . A process that solves such an equation is usually referred to as a *diffusion process*. If  $X$  and  $b$  take values in  $\mathbb{R}^d$ ,  $\sigma$  is a  $\mathbb{R}^{d \times d'}$ -valued matrix for  $d' \leq d$  and  $W$  an  $\mathbb{R}^{d'}$ -valued Brownian motion, equation (2.27) is understood to represent the system of SDEs for the vector  $X$  with entries  $X^i$  that solve

$$dX_t^i = b^i(t, X_t) dt + \sum_{j=1}^{d'} \sigma_{ij}(t, X_t) dW_t^j, \quad i = 1, \dots, d.$$

For the remainder of this thesis, we will omit the summation in the right hand side and apply Einstein's summation convention from 2.1.16.

A question that one can ask, is whether or not an SDE, given a drift and diffusion coefficient has a solution and if so, in what sense that solution may or may not be unique. We therefore pose the following theorem adapted from theorem 5.2.1 of Øksendal, 2003 [17].

**Theorem 2.2.9** (Existence and uniqueness theorem for SDEs). *Let  $T > 0$  and suppose that  $b : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$  and  $\sigma : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d'}$  be measurable functions that satisfy global Lipschitz conditions given by*

$$|b(t, x)| + |\sigma(t, x)| \leq C(1 + |x|), \quad x \in \mathbb{R}^d, t \in [0, T] \quad (2.28)$$

for some  $C > 0$ , where  $|\sigma|^2 = \sum_{i,j} |\sigma_{ij}|^2$ , and

$$|b(t, x) - b(t, y)| + |\sigma(t, x) - \sigma(t, y)| \leq D|x - y|, \quad x, y \in \mathbb{R}^d, t \in [0, T] \quad (2.29)$$

for some  $D > 0$ . Let  $W$  be a standard Brownian motion in  $\mathbb{R}^{d'}$  and let  $\mathcal{F}_t = \sigma(W_s : s \leq t)$  denote the  $\sigma$ -field generated by  $W$ . Furthermore let  $Z$  be a random variable in  $\mathbb{R}^d$  independent of  $\mathcal{F}_\infty$  such that  $\mathbb{E}(|Z|^2) < \infty$ . Then the stochastic differential equation

$$dX_t = b(t, X_t) dt + \sigma(t, X_t) dW_t, \quad t \in [0, T], X_0 = Z$$

has a unique solution  $X$  that is continuous in  $t$ , adapted to  $(\mathcal{F}_t)_t$  and

$$\mathbb{E} \int_0^T |X_t|^2 dt < \infty.$$

Here uniqueness is interpreted as follows. There is a unique measurable map  $X$  from  $\Omega$  to the space of functions  $[0, T] \rightarrow \mathbb{R}^d$  that maps  $\mathbb{P}$ -almost each  $\omega \in \Omega$  to a function  $X(\omega)$  that solves the equation for that particular path  $W(\omega)$ .

*Proof.* See Øksendal, 2003 [17]. □

A well-known result in the field of stochastic differential equations is Itô's formula.

**Theorem 2.2.10** (Itô's lemma). *Let  $f \in \mathcal{C}^2(\mathbb{R}^d)$  and let  $X$  be a stochastic process that takes values in  $\mathbb{R}^d$ . Then the Itô formula holds*

$$df(X_t) = \partial_i f(X_t) dX_t^i + \frac{1}{2} \partial_i \partial_j f(X_t) d[X^i, X^j]_t \quad (2.30)$$

where we have used Einsteins summation convention from 2.1.16 to sum over  $i$  and  $j$  and, similar to the standard notation on manifolds, we set  $\partial_i = \frac{\partial}{\partial x^i}$ .

*Proof.* For the proof of this theorem, we refer to section 4.2 of Øksendal, 2003 [17].  $\square$

Another relation between the Itô integral and the quadratic covariation is given in the following theorem.

**Theorem 2.2.11.** *Let  $X$  and  $Y$  be continuous martingales that are bounded in  $L^2$ . Then  $X \cdot Y$  is the unique martingale that satisfies  $[X \cdot Y, Z] = X \cdot [Y, Z]$  for all continuous martingales  $Z$  that are  $L^2$ -bounded.*

*Proof.* This theorem is part of theorem 5.4 of Le Gall, [12].  $\square$

It is a well-known result that if  $W$  is a Brownian motion of dimension 1, that  $[W, W]_t = t$  (See A.3.9) and that for  $f \in \mathcal{C}^2(\mathbb{R})$ , the Itô formula reads

$$df(W_t) = \frac{1}{2} f''(W_t) dt + f'(W_t) dW_t. \quad (2.31)$$

Another important corollary of the Itô formula is the product rule for stochastic processes

**Corollary 2.2.12.** *For two continuous semimartingales  $X$  and  $Y$ , we have*

$$dX_t Y_t = X_t dY_t + Y_t dX_t + d[X, Y]_t \quad (2.32)$$

Now notice that the equations (2.31) and (2.32) differ from the usual rules of calculus where  $df(x) = f'(x) dx$ . To resolve this problem, the Stratonovich form of stochastic differential equations is introduced

**Definition 2.2.13** (Stratonovich integral). *Under the same assumptions as in definition 2.2.7, we define the Stratonovich integral as the process*

$$(H \circ X)_t = \int_0^t H_s \circ dX_s = \int_0^t H_s dX_s + \frac{1}{2} [H, X]_t \quad (2.33)$$

The stochastic differential equations associated with the Stratonovich integral are usually written down as

$$dX_t = b(t, X_t) dt + \sigma(t, X_t) \circ dW_t \quad (2.34)$$

**Theorem 2.2.14** (Stratonovich to Itô conversion). *Equation (2.34) coincides with the Itô equation*

$$dX_t^i = \left( b^i(t, X_t) + \frac{1}{2} \sum_{j,k} \sigma_{jk}(t, X_t) \frac{\partial \sigma_{ik}}{\partial x^j}(t, X_t) \right) dt + \sigma_{ij}(t, X_t) dW_t^j$$

*Proof.* We will show this result for an  $\mathbb{R}$ -valued process  $X$ , as the proof in higher dimensions follows in a similar fashion, but with more complicated equations. Notice that, a solution  $X$  to (2.34) can be written as

$$X_t = X_0 + \int_0^t b(s, X_s) ds + \int_0^t \sigma(s, X_s) dW_s + \frac{1}{2} [\sigma, W]_t$$

It now follows that

$$\begin{aligned} [\sigma, W] &= \lim_{\|\Delta\| \rightarrow 0} \sum_{i=1}^N (\sigma(t_i, X_{t_i}) - \sigma(t_{i-1}, X_{t_{i-1}})) (W_{t_i} - W_{t_{i-1}}) \\ &= \lim_{\|\Delta\| \rightarrow 0} \sum_{i=1}^N \frac{\sigma(t_i, X_{t_i}) - \sigma(t_{i-1}, X_{t_{i-1}})}{X_{t_i} - X_{t_{i-1}}} (W_{t_i} - W_{t_{i-1}}) (X_{t_i} - X_{t_{i-1}}) \\ &= \lim_{\|\Delta\| \rightarrow 0} \sum_{i=1}^N \frac{\Delta \sigma(t_i, X_{t_i})}{\Delta X_{t_i}} \Delta W_{t_i} \Delta X_{t_i} \end{aligned}$$

Now notice that when taking this limit, continuity of  $X$  implies that the first term converges to  $\frac{\partial \sigma}{\partial x}(t_i, X_{t_i})$ , whereas the second product converges to  $dW_t dX_t = b(t, X_t) dt dW_t + \sigma(t, X_t) dW_t dW_t + \frac{1}{2} d[\sigma, W] dW_t$ . Note that, when taking expectations, only the term in the middle remains and  $\mathbb{E}(dW dW) = dt$  and thus we can arrive at

$$[\sigma, W] = \lim_{\|\Delta\| \rightarrow 0} \sum_{i=1}^N \frac{\Delta \sigma(t_i, X_{t_i})}{\Delta X_{t_i}} \sigma(t_i, X_{t_i}) \Delta t = \int_0^t \frac{\partial \sigma}{\partial x}(s, X_s) \sigma(s, X_s) ds$$

We thus arrive at

$$X_t = X_0 + \int_0^t \left( b(s, X_s) + \frac{1}{2} \frac{\partial \sigma}{\partial x}(s, X_s) \sigma(s, X_s) \right) ds + \int_0^t \sigma(s, X_s) dW_s + \frac{1}{2} [\sigma, W]_t$$

which concludes the proof.  $\square$

As stated earlier, the Stratonovich integral is introduced, because it behaves like the Lebesgue integral. Let us demonstrate this behavior in the following proposition.

**Proposition 2.2.15** (Stratonovich calculus). *If  $f \in \mathcal{C}^3(\mathbb{R}^d)$ , then the change of variable formula holds.*

$$df(X_t) = \partial_i f(X_t) \circ dX_t^i.$$

One thus has that if  $F: \mathbb{R}^d \rightarrow \mathbb{R}^n$  is of class  $\mathcal{C}^3$  and  $X$  denotes a  $d$ -dimensional process, that

$$dF(X_t) = J_F(X_t) \circ dX_t,$$

where  $J_F$  denotes the Jacobian matrix of  $F$  with entries  $\partial_j F^i$ .

Notice that in integral form, the result does match the change of variable formula as we know it from ordinary calculus as it then reads

$$f(X_t) = f(X_0) + \int_0^t \partial_i f(X_s) \circ dX_s^i$$

*Proof.* Let us go back to the Itô formula (2.30). If we substitute the definition of the Stratonovich integral into it, notice that it reads

$$df(X_t) = \partial_i f(X_t) \circ dX_t^i - \frac{1}{2} d[\partial_i f(X), X^i]_t + \frac{1}{2} \partial_i \partial_j f(X_t) d[X^i, X^j]_t \quad (2.35)$$

Since  $f$  is of class  $\mathcal{C}^3$ , we can now apply the Itô formula to  $\partial_i f(X)$  to find that

$$d \partial_i f(X_t) = \partial_k \partial_i f(X_t) dX_t^k + \frac{1}{2} \partial_k \partial_j \partial_i f(X_t) d[X^k, X^j]_t$$

When integrating both sides, we find that

$$\partial_i f(X_t) = \partial_i f(X_0) + (\partial_k \partial_i f(X) \cdot X^k)_t + \left( \frac{1}{2} \partial_k \partial_j \partial_i f(X) \cdot [X^k, X^j] \right)_t$$

Using the calculation rules of quadratic variation then gives us that

$$[\partial_i f(X), X^i]_t = [\partial_k \partial_i f(X) \cdot X^k, X^i]_t + \frac{1}{2} [\partial_k \partial_j \partial_i f(X) \cdot [X^k, X^j], X^i]_t$$

In A.3.6 we will show that the quadratic covariation of two semimartingales is of finite total variation and we will follow up on that by showing that if a process has finite total variation, the quadratic variation with another semimartingale is 0 in A.3.7. Since  $X$  solves an SDE, it is a continuous semimartingale and therefore the latter term in our last equation will be 0. We can thus derive that

$$d[\partial_i f(X), X^i]_t = \partial_k \partial_i f d[X^k, X^i]_t$$

The desired result now follows by smartly rearranging the indices.  $\square$

### 2.2.3. Transition kernels, Kolmogorov equations and the infinitesimal generator

In this section, we only consider time-homogeneous diffusions, that is  $b$  and  $\sigma$  do not have a direct dependence on time. Let  $X = (X_t)_t$  be a time-homogeneous diffusion process that takes values in a Banach space  $(\mathcal{X}, \|\cdot\|)$  where  $\mathcal{B}$  denotes the Borel  $\sigma$ -field generated by the open sets. The norm of functions is understood to be the supremum norm  $\|f\| = \sup_{x \in \mathcal{X}} |f(x)|$ . We denote its transition kernel by  $(P_t)_t$ , that is

$$P_t(x, A) = \mathbb{P}(X_t \in A \mid X_0 = x), \quad x \in \mathcal{X}, A \in \mathcal{B}$$

Now if  $X$  is a continuous diffusion process,  $(P_t)_t$  satisfies the following properties:

- $P_t(\cdot, A)$  is a measurable function  $\mathcal{X} \rightarrow [0, 1]$  for all  $A \in \mathcal{B}$  and  $t \geq 0$ .
- $P_t(x, \cdot)$  is a probability measure on  $(\mathcal{X}, \mathcal{B})$  for all  $x \in \mathcal{X}$  and  $t \geq 0$ .

For any  $t \geq 0$  and  $x \in \mathcal{X}$ , this leads us to an operator on the set of all measurable functions  $\mathcal{X} \rightarrow \mathbb{R}$  given by

$$(P_t f)(x) = \mathbb{E}^x f(X_t) = \int_{\mathcal{X}} f(y) P_t(x, \mathbf{d}y)$$

where the superscript indicates that the expectation is taken conditional on  $X_0 = x$ . Note that this operator generalizes the kernel since one still has that  $P_t(x, A) = (P_t \mathbb{1}_A)(x)$ .

An important property of diffusion processes, is that they are continuous Markov processes (see 7.1.2 in Øksendal, 2003 [17]). That is, they satisfy the Markov property.

**Definition 2.2.16** (Markov Processes). *An  $(\mathcal{F}_t)_t$ -adapted stochastic process  $X = (X_t)_t$  is a Markov process with initial distribution  $\nu$  if  $\nu$  is a probability measure on  $(\mathcal{X}, \mathcal{B})$  and the following criteria are met.*

- (i)  $\mathbb{P}(X_0 \in B) = \nu(B)$  for all  $B \in \mathcal{B}$ .
- (ii)  $\mathbb{E}^\nu(f(X_{t+s}) \mid \mathcal{F}_s) = P_t f(X_s)$  for all measurable  $f : \mathcal{X} \rightarrow \mathbb{R}$  and  $s, t \geq 0$ .

**Theorem 2.2.17.** *The transition kernel  $\{P_t\}_t$  of an  $(\mathcal{X}, \mathcal{B})$ -valued almost surely continuous process  $X = (X_t)_t$  forms a strongly continuous semigroup on the space of measurable functions  $\mathcal{X} \rightarrow \mathbb{R}$ , i.e.  $\{P_t\}_t$  satisfies the following conditions.*

- (i)  $P_t$  is a bounded linear operator for all  $t \geq 0$ , that is  $P_t$  is linear and  $\|P_t\| = \sup_{\|f\|=1} \|P_t f\| < \infty$ .
- (ii)  $\lim_{t \downarrow 0} \|P_t f - f\| = 0$  for all measurable  $f : \mathcal{X} \rightarrow \mathbb{R}$ .

*Proof.* By linearity of expectations,  $P_t$  is a linear operator and since  $P_t(x, \cdot)$  is a probability measure on  $\mathcal{X}$  for all  $x \in \mathcal{X}$ , we also have that

$$\|P_t\| = \sup_{\|f\|=1} \|P_t f\| = \sup_{\|f\|=1} \sup_{x \in \mathcal{X}} \left| \int_{\mathcal{X}} f(y) P_t(x, \mathbf{d}y) \right| \leq \sup_{\|f\|=1} \sup_{x \in \mathcal{X}} \int_{\mathcal{X}} |f(y)| P_t(x, \mathbf{d}y) \leq \sup_{x \in \mathcal{X}} \int_{\mathcal{X}} P_t(x, \mathbf{d}y) = 1$$

and thus  $P_t$  is indeed a bounded linear operator for all  $t$ . For the second condition, it is important to notice that for any  $x \in \mathcal{X}$ , we have

$$|P_t f(x) - f(x)| \leq \int_{\mathcal{X}} |f(y) - f(x)| P_t(x, \mathbf{d}y) = \int_{\mathcal{X}} |f(y) - f(x)| \mathbb{P}(X_t \in \mathbf{d}y \mid X_0 = x)$$

Now whenever  $y \neq x$ , by continuity of the process  $X$ , the events  $\{X_t \in \mathbf{d}y\}$  shrink to a null set as  $t \downarrow 0$  conditional on  $X_0 = x$ . Therefore, by monotonicity of measures, we must have that

$$\lim_{t \downarrow 0} |P_t f(x) - f(x)| = 0$$

for all  $x \in \mathcal{X}$  and thus  $\|P_t f - f\| \rightarrow 0$  whenever  $t \downarrow 0$ . □

**Remark 2.2.18.** *In this proof we introduced the norm of a bounded linear operator in a similar manner as in Bressan, 2012 [4]. An important observation is that for a measurable function  $f : \mathcal{X} \rightarrow \mathbb{R}$  this norm satisfies*

$$\|P_t f\| = \left\| P_t \left( \frac{f}{\|f\|} \right) \right\| \|f\| \leq \|P_t\| \|f\| \tag{2.36}$$

The second identity of the definition of Markov processes now gives us that

$$P_t(P_s f)(x) = \mathbb{E}^x P_s f(X_t) = \mathbb{E}^x \mathbb{E}^\nu(f(X_{t+s}) | \mathcal{F}_t) = \mathbb{E}^x f(X_{t+s}) = P_{t+s} f(x)$$

and therefore  $P_{t+s} = P_t P_s$ . For the second but last equality, we refer to theorem A.1.2 in appendix A. When this result is combined with theorem 2.2.17, we observe that the map  $t \mapsto P_t f$  is continuous for any measurable function  $f : \mathcal{X} \rightarrow \mathbb{R}$ , since equation (2.36), combined with linearity of  $P_t$  and theorem 2.2.17, implies that  $\|P_{t+h} f - P_t f\| \leq \|P_t\| \|P_h f - f\| \rightarrow 0$  as  $h \downarrow 0$  (left continuity follows similarly). For convenience, we will denote  $P_{t,s}(x, A) := \mathbb{P}(X_t \in A | X_s = x)$  for  $t < s$ , and then the Markov property gives us that  $P_{t,s} = P_{t-s}$

**Definition 2.2.19** (Transition density). *The transition density  $p$  of  $X$  is defined as the Radon-Nikodym derivative of  $P_{t,s}$  with respect to the Lebesgue measure, provided it exists. That is,  $\mathbb{P}(X_t \in dy | X_s = x) = p(s, x; t, y) dy$  for  $s < t$ . We frequently take  $t = T$  fixed and abbreviate  $p$  for convenience as  $p(s, x)$ .*

**Example 2.2.20.** *From equation (2.23), it follows that the transition kernel for a standard Brownian motion  $W$  starting at  $x$  is given by*

$$(P_t f)(x) = \mathbb{E}^x f(W_t) = \int \frac{1}{\sqrt{2\pi t}} e^{-\frac{(y-x)^2}{2t}} f(y) dy$$

and we deduce that

$$p(s, x; t, y) = \frac{1}{\sqrt{2\pi(t-s)}} e^{-\frac{(y-x)^2}{2(t-s)}}$$

In the field of (functional) analysis, partial differential equations have been studied that are characterized by a differential operator. For example, the heat equation is characterized by the Laplace operator. In the field of stochastic differential equations, we can see something similar, as stochastic processes can be described on small time intervals via their *infinitesimal generator* as well. This is a partial differential operator that characterizes the transition kernel of stochastic processes. We will now set the domain for the desired operator to be

$$\mathcal{D} := \left\{ f : \mathcal{X} \rightarrow \mathbb{R} \text{ measurable} : \text{There exists a measurable } g : \mathcal{X} \rightarrow \mathbb{R} \text{ such that } \lim_{t \downarrow 0} \left\| \frac{P_t f - f}{t} - g \right\| = 0 \right\}.$$

Here, the norm is still understood to be the uniform norm  $\|f\| = \sup_{x \in \mathcal{X}} |f(x)|$ .

**Definition 2.2.21** (Infinitesimal generator). *On  $\mathcal{D}$ , the infinitesimal generator  $\mathcal{L}$  for is defined as*

$$\mathcal{L} f := \lim_{t \downarrow 0} \frac{P_t f - f}{t}$$

The domain  $\mathcal{D}$  of  $\mathcal{L}$  is typically denoted by  $\mathcal{D}(\mathcal{L})$ .

It is important to notice that  $\mathcal{D}(\mathcal{L})$  need not be non-empty. However, if  $\mathcal{D}(\mathcal{L}) \neq \emptyset$ , the definition immediately leads to the the identity

$$\mathbb{E}^\nu(f(X_{t+h}) - f(X_t) | \mathcal{F}_t) = h\mathcal{L}f(X_t) + o(h), \quad \mathbb{P}_\nu - \text{a.s.},$$

as  $h \downarrow 0$  for all  $f \in \mathcal{D}(\mathcal{L})$ . In this sense, the infinitesimal generator does describe the motion of our process in an infinitesimal time-interval.

**Remark 2.2.22.** *In chapter III.5 of Rogers and Williams, 2000 [20], it is derived from the Hille-Yoshida theorem that the tuple  $(\mathcal{L}, \mathcal{D}(\mathcal{L}))$  uniquely determines  $\{P_t\}$ .*

**Definition 2.2.23** ((Uniform) ellipticity). *We call the diffusion  $X$  (uniformly) elliptic if the infinitesimal generator is (uniformly) elliptic. That is, the matrix  $(a_{ij}(x))_{i,j}$  is strictly positive definite for all  $x$  (elliptic) and all eigenvalues are  $\geq \theta$  for some uniform  $\theta > 0$  for all  $x$  (uniformly elliptic).*

**Theorem 2.2.24** (Infinitesimal generator of diffusion processes). *Let  $X$  be  $d$ -dimensional diffusion process with drift  $b$  and diffusion coefficient  $\sigma$ . Then we have that the infinitesimal generator has domain  $\mathcal{D}(\mathcal{L}) := \mathcal{C}_0^2(\mathbb{R}^d)$ , where  $\mathcal{C}_0^2(\mathbb{R}^d)$  denotes the set of all functions of class  $\mathcal{C}^2$  that have compact support. Furthermore, the infinitesimal generator is given in closed form by*

$$\mathcal{L} = b^i \partial_i + \frac{1}{2} a^{ij} \partial_i \partial_j. \quad (2.37)$$

where  $a$  denotes the matrix  $a = \sigma \sigma^T$ .

*Proof.* This is a direct consequence of lemma 7.3.2 of Øksendal, 2003 [17] where one can take stopping times  $\tau = t$  and apply that  $P_t f(x) = \mathbb{E}^x f(X_t)$ .  $\square$

**Corollary 2.2.25.** *For a Brownian motion  $W$  in  $\mathbb{R}^d$ , we have that  $\mathcal{D}(\mathcal{L}) = \mathcal{C}_0^2(\mathbb{R}^d)$ . Furthermore, for  $f \in \mathcal{D}(\mathcal{L})$ , we have  $\mathcal{L} = \frac{1}{2} \Delta$ , where  $\Delta$  denotes the Laplace operator  $\Delta = \sum_{i=1}^d \partial_i^2$ .*

**Theorem 2.2.26** (Kolmogorov equations). *Let  $\{P_t\}_t$  be the transition kernel of a time homogeneous diffusion process taking values in a Banach space  $\mathcal{X}$  with infinitesimal generator  $\mathcal{L}$  and let  $f \in \mathcal{D}(\mathcal{L})$  and  $t \geq 0$ . Then  $P_t f \in \mathcal{D}(\mathcal{L})$ . The function  $t \mapsto P_t f$  is differentiable and the Kolmogorov backward and forward equations hold:*

$$\frac{d}{dt} P_t f \stackrel{\text{(backward)}}{=} \mathcal{L} P_t f \stackrel{\text{(forward)}}{=} P_t \mathcal{L} f$$

*Proof.* Since  $X$  is a time homogeneous diffusion process, it is also a Markov process and therefore

$$\left\| \frac{P_{t+h} f - P_t f}{h} - P_t \mathcal{L} f \right\| = \left\| P_t \left( \frac{P_h f - f}{h} - \mathcal{L} f \right) \right\| \leq \|P_t\| \left\| \frac{P_h f - f}{h} - \mathcal{L} f \right\|.$$

Similarly, we also have

$$\begin{aligned} \left\| \frac{P_t f - P_{t-h} f}{h} - P_t \mathcal{L} f \right\| &\leq \left\| \frac{P_t f - P_{t-h} f}{h} - P_{t-h} \mathcal{L} f \right\| + \|P_{t-h} \mathcal{L} f - P_t \mathcal{L} f\| \\ &\leq \|P_{t-h}\| \left\| \frac{P_h f - f}{h} - \mathcal{L} f \right\| + \|P_{t-h} \mathcal{L} f - P_t \mathcal{L} f\| \end{aligned}$$

Since  $P_t$  is continuous in  $t$ , taking the limit  $h \downarrow 0$  yields

$$\lim_{h \downarrow 0} \frac{P_{t+h} f - P_t f}{h} = \lim_{h \downarrow 0} \frac{P_t f - P_{t-h} f}{h} = P_t \mathcal{L} f$$

We have thus proven differentiability as well as the Kolmogorov forward equation. Now observe that, when using that  $P_{t+s} = P_t P_s$ , these equations directly imply that indeed  $P_t f \in \mathcal{D}(\mathcal{L})$  and that the backward equation follows from the definition of  $\mathcal{L}$ .  $\square$

## 2.3. The Transition Density

We continue to let  $X$  be a time homogeneous process. In the previous section, we briefly introduced the transition density  $p$  of a stochastic processes the density of the transition kernel of the process,  $\mathbb{P}(X_t \in dy \mid X_s = x) = p(s, x; t, y) dy$ . In this section we further elaborate on the transition density.

Let us briefly return to the problem as formulated in the introduction. We consider a vector of data points  $\xi = (\xi_1, \dots, \xi_n)$  in  $\mathbb{R}^d$  modeled as realizations of  $X_T$ . The likelihood of  $\theta$  is given by

$$L(\theta \mid \xi) = \prod_{i=1}^n p(0, \theta; T, \xi_i) \quad (2.38)$$

This result shows that when  $p$  is known, we can derive a distribution for the center of our data, so we thus need to find out when  $p$  exists in closed form and find a way approximate  $p$ .

### 2.3.1. Finding the transition density of a diffusion process

Assume  $X$  is a time-homogeneous diffusion process characterized by the stochastic differential equation

$$dX_t = b(X_t) dt + \sigma(X_t) dW_t, \quad X_0 = \theta$$

Since likelihood-based inference is hampered by the fact that the transition density is not known in closed form, we rephrase the Kolmogorov equations in terms of the transition density. Let  $P_{t,s}$  denote the transition kernel of  $X$ , i.e.  $P_{t,s}(x, A) = \mathbb{P}(X_t \in A \mid X_s = x)$ . Set  $t = T - s$  and suppose that  $P_{t,s}$  admits a density  $p$  with respect to the Lebesgue measure. Since diffusion processes are Markov processes, we have that  $P_t(x, dy) = p(s, x; T, y) dy$ .

Recall from theorem 2.2.24 that the process  $X$  has an infinitesimal generator given by

$$\mathcal{L} = b^i \partial_i + \frac{1}{2} a^{ij} \partial_i \partial_j.$$

where  $a(x) = \sigma(x) \sigma(x)^T$ . Recall from theorem 2.2.26 that the Kolmogorov backward equation states that

$$\frac{d}{dt} P_t(x, A) = (\mathcal{L} P_t(\cdot, A))(x)$$

Now since  $t = T - s$ , we have  $dt = -ds$ , it follows that

$$-\frac{\partial}{\partial s} \int_A p(s, x; T, y) dy = \left( \mathcal{L} \int_A p(s, \cdot; T, y) dy \right)(x)$$

Since  $p$  is a probability density and  $\mathcal{L}$  is a differential operator while the integral on the right hand side does not depend on  $x$ , we may again rewrite this equation as

$$-\int_A \frac{\partial}{\partial s} p(s, x; T, y) dy = \int_A (\mathcal{L} p(s, \cdot; T, y))(x) dy$$

Since this integral equation has to hold for any Borel set  $A$ , we arrive at

$$-\frac{\partial p}{\partial s}(s, x) = b^i(s, x) \partial_i p(s, x) + \frac{1}{2} a^{ij}(s, x) \partial_i \partial_j p(s, x) \quad (2.39)$$

where  $p(s, x)$  is short-hand notation for  $p(s, x; T, y)$ . Equation (2.39) is an equation that appears a lot in the theory of stochastic differential equations and is also known as the *Fokker-Planck equation*.

At first glance, it is not clear whether or not equation (2.39) has a solution. On closer notice however, we can see that the partial differential equation is linear in the partial derivatives of  $p$  and thus we can use the results seen in the following section from functional analysis. We now move to a small section on parabolic equations in order to derive conditions such that equation (2.39) indeed has a solution.

### 2.3.2. Parabolic equations

The theory of parabolic equations in functional analysis is build to investigate whether or not equations of the same type as equation (2.39) have a solution. We base this section on the theory seen in Bressan, 2012 [4].

Let us start with formally defining the problem. Let  $\Omega \subseteq \mathbb{R}^d$  be a bounded open set and, given measurable functions  $a^{ij}, b^i, c : \Omega \rightarrow \mathbb{R}$ , consider the second order differential operator

$$\mathcal{L}u = -\partial_j (a^{ij} \partial_i u) + \partial_i (b^i u) + cu \quad (2.40)$$

Consider the initial value problem

$$\begin{cases} \frac{\partial u}{\partial t} + \mathcal{L}u &= 0, & t > 0, x \in \Omega, \\ u(t, x) &= 0, & t > 0, x \in \partial\Omega \\ u(0, x) &= g(x), & x \in \Omega \end{cases} \quad (2.41)$$

In order to investigate whether or not there is a solution to this initial problem we will use the following assumption throughout this section.

**Assumption 2.3.1.** *The domain  $\Omega \subseteq \mathbb{R}^d$  is open and bounded. The coefficients  $a^{ij}, b^i$  and  $c$  in equation (2.40) are in  $L^\infty(\Omega)$ . Moreover, the operator  $\mathcal{L}$  is uniformly elliptic, i.e. there exists a constant  $\theta > 0$  such that*

$$a^{ij}(x) \xi_i \xi_j \geq \theta \|\xi\|^2 \quad \text{for all } x \in \Omega \text{ and } \xi \in \mathbb{R}^d \quad (2.42)$$

Assumption 2.3.1 basically means that for any  $x$ , the matrix with coefficients  $a^{ij}(x)$  must be strictly positive definite.

**Remark 2.3.2.** *Notice that differential operators of the form*

$$\mathcal{L}u = -a^{ij} \partial_i \partial_j u + b^i \partial_i u + cu$$

can be rewritten in the form of equation (2.40) via

$$\mathcal{L}u = -\partial_j (a^{ij} \partial_i u) + \partial_i ((\partial_j a^{ij} + b^i) u) + (c - (\partial_i \partial_j a^{ij} - \partial_i b^i)) u$$

and we can thus treat these problems, under the condition that all  $a^{ij}, \partial_i \partial_j a^{ij}, b^i, \partial_i b^i, c \in L^\infty(\Omega)$ , using the same theory we encounter in this section.

Since we are in the scenario of remark 2.3.2, we have to make one assumption in addition to assumption 2.3.1.

**Assumption 2.3.3.** *For all  $i, j = 1, \dots, d$ ,  $a^{ij}$  is an element of the Sobolev space  $W^{2,\infty}(\Omega)$  consisting of all  $L^\infty$  functions  $\Omega \rightarrow \mathbb{R}$  that have their first two weak derivatives in  $L^\infty(\Omega)$  and  $b^i$  is an element of  $W^{1,\infty}(\Omega)$  and thus has its first weak derivative in  $L^\infty(\Omega)$ .*

**Definition 2.3.4** (Weak solutions). *A function  $u \in H_0^1(\Omega)$  that satisfies*

$$\int_{\Omega} (\mathcal{L}u) v \, dx = \int_{\Omega} f v \, dx \quad \text{for all } v \in \mathcal{C}_0^\infty(\Omega)$$

is called a weak solution to (2.41). Here  $\mathcal{C}_c^\infty(\Omega)$  represents the set of all  $\mathcal{C}^\infty$ -functions  $\Omega \rightarrow \mathbb{R}$  with compact support and  $H_0^1(\Omega) = W_0^{1,2}(\Omega)$  represents the Hilbert-Sobolev space of all  $L^2$ -functions  $\Omega \rightarrow \mathbb{R}$  that vanish at  $\partial\Omega$  and are such that their first weak derivatives are in  $L^2(\Omega)$  as well.

Now a convenient way to look at weak solutions is to consider them as bilinear forms. So on the Hilbert-Sobolev space  $H_0^1(\Omega)$ , consider the bilinear form

$$B(u, v) := \int_{\Omega} (\mathcal{L}u) v \, dx \quad (2.43)$$

Then we can consider  $u$  to be a weak solution to (2.41) if

$$B(u, v) = \langle u, v \rangle_{L^2} \quad \text{for all } v \in \mathcal{C}_c^\infty(\Omega)$$

where  $\langle \cdot, \cdot \rangle_{L^2}$  denotes the standard inner product on  $L^2$  given by  $\langle u, v \rangle_{L^2} = \int u(x) v(x) \, dx$

**Remark 2.3.5.** *Via integration by parts, one can show that*

$$\int_{\Omega} (\mathcal{L}u) v \, dx = \int_{\Omega} [a^{ij} (\partial_i u) (\partial_j v) - b^i u (\partial)_i v + cuv] \, dx \quad (2.44)$$

**Theorem 2.3.6** (Semigroup of solutions of a parabolic equation). *Suppose assumptions 2.3.1 and 2.3.3 hold. Moreover, suppose that the bilinear form defined in equation (2.43) is strictly positive definite. Then the operator  $-\mathcal{L}$  generates a contractive semigroup  $\{S_t : t \geq 0\}$  of linear operators on  $L^2(\Omega)$ . That is,  $-\mathcal{L}$  satisfies the following:*

- Each  $S_t$  as a bounded linear operator.
- For each pair  $t, s \geq 0$ , the composition satisfies  $S_t S_s = S_{t+s}$  and  $S_0 = I$ .
- For every  $u \in L^2(\Omega)$ , the map  $t \mapsto S_t u$  is continuous.

- $\|S_t\| \leq 1$  and  $\lim_{t \downarrow 0} \frac{S_t u - u}{t} = -\mathcal{L}u$  for each  $u$  for which this limit exists.

*Proof.* For this proof, we refer to theorem 9.21 of Bressan, 2012 [4]. □

Theorem 2.3.6 now provides us with all that we needed in this chapter, as the map  $t \mapsto u(t) := S_t g$  is a continuous map  $[0, \infty) \rightarrow L^2(\Omega)$  and satisfies  $u(0) = g$ . The initial conditions in (2.41) are thus satisfied. Moreover, if  $g \in \mathcal{D}(-\mathcal{L})$ , then so is  $S_t g$  for all  $t \geq 0$  and theorem 7.6 of Bressan, 2012 [4] even demonstrates the map is continuously differentiable and since  $\mathcal{D}(-\mathcal{L}) \subseteq H_0^1(\Omega)$ , the boundary conditions in (2.41) are also satisfied. This shows that the map  $(t, x) \mapsto S_t g(x)$  solves (2.41).

This theorem concludes the section on parabolic equation as we have proven that under certain assumptions, the partial differential equation for the transition density can indeed be solved with these boundary conditions. In chapter 9 of Bressan, 2012 [4], this result is extended to different boundary conditions.

### 2.3.3. Conditions for the existence of a transition density

In the start of this section on transition densities, we have derived a parabolic equation (2.39) for the transition density of a diffusion process. After that, we went into a bit of functional analysis in section 2.3.2 to show that, under certain conditions, parabolic equations can be solved. Going back to the equation  $dX_t = b(X_t) dt + \sigma(X_t) dW_t$ , we note that  $a = \sigma\sigma^T$  is a positive definite matrix and therefore satisfies assumption 2.3.1. We have thus derived that a transition density for  $X_t$  indeed exists as long as  $a$  and  $b$  have their derivatives in  $L^\infty$ . These assumptions, however, are very reasonable assumptions as we can see in theorem 2.2.9, that global Lipschitz conditions can also be imposed to ensure that the stochastic differential equations have a unique solution. This motivates the following requirements for the SDEs.

$$|b(t, x)| + |\sigma(t, x)| \leq C(1 + |x|) \quad (2.45)$$

$$|b(t, x) - b(t, y)| + |\sigma(t, x) - \sigma(t, y)| \leq D|x - y| \quad (2.46)$$

**Example 2.3.7.** One can quickly verify that a  $\mathcal{C}^2$ -function  $\sigma$ , generates a Stratonovich SDE  $dX_t = \sigma(X_t) \circ dW_t$  satisfies equations (2.45) and (2.46) and it thus has a unique solution that admits a transition density.

Notice that these conditions guarantee the existence and uniqueness of a solution of an SDE and a transition density, but that this is no equivalence, as a unique solution with a density can still exist when these assumptions are not met. For example, consider the system of SDEs given by

$$\begin{aligned} dX_t &= Y_t dt \\ dY_t &= b(Y_t) dt + \gamma(Y_t) dW_t \end{aligned}$$

where  $b$  and  $\gamma$  are smooth functions. The diffusion coefficient matrix of this system is given by

$$\sigma(x, y) = \begin{pmatrix} 0 \\ \gamma(x) \end{pmatrix}, \quad \text{and} \quad a(x, y) = \begin{pmatrix} 0 & 0 \\ 0 & \gamma(y) \end{pmatrix}.$$

Clearly,  $Y$  has a unique solution as  $b$  and  $\gamma$  are smooth and thus  $X$  has a unique solutions as well. However the matrix  $a$  is singular and thus has eigenvalue 0 and therefore assumption 2.3.1 is not satisfied.

# 3

## Brownian Motion on a Riemannian manifold

With the background theory developed in chapter 2, we can now characterize Brownian motion on Riemannian manifolds. This chapter introduces two characterizations of Brownian motion on a Riemannian manifold  $(\mathcal{M}, g)$ .

### 3.1. Brownian motion in local coordinates

In this section we introduce a first characterization of Brownian motion on a Riemannian manifold that is obtained using the Laplace-Beltrami operator.

#### 3.1.1. Derivation of the SDE for Brownian motion in local coordinates

Recall from remark 2.2.22 that under weak conditions, a diffusion process can be defined via its generator. Let  $X$  be the diffusion process governed by the SDE

$$dX_t^i = b^i(X_t) dt + \sigma_{ij}(X_t) dW_t^j, \quad i = 1, \dots, d \quad (3.1)$$

where  $(W_t)_{t \geq 0}$  is an  $\mathbb{R}^{d'}$ -valued Wiener process with  $d' \leq d$ . Then  $X$  has infinitesimal generator

$$(\mathcal{L}f)(x) = \frac{1}{2} a^{ij}(x) \partial_i \partial_j f(x) + b^i(x) \partial_i f(x) \quad (3.2)$$

where  $a^{ij}(x) = (\sigma(x) \sigma(x)^T)_{ij}$ . Now let us revisit equation (2.21) describing the Laplace-Beltrami operator as

$$\Delta_{\mathcal{M}} f = g^{ij} \partial_i \partial_j f - g^{jk} \Gamma_{jk}^i \partial_i f.$$

Recall that in section 2.2.3 we derived that a Brownian motion in  $\mathbb{R}^d$  is generated by  $\frac{1}{2} \Delta$ . It therefore makes sense to define Brownian motion on a Riemannian manifold  $(\mathcal{M}, g)$  as the process generated by  $\frac{1}{2} \Delta_{\mathcal{M}}$ . We thus have that, in local coordinates

$$\begin{aligned} a^{ij}(x) &= g^{ij}(x), & i, j &= 1, \dots, d \\ b^i(x) &= -\frac{1}{2} g^{jk}(x) \Gamma_{jk}^i(x), & i &= 1, \dots, d \end{aligned} \quad (3.3)$$

Since  $(g^{ij})$  is symmetric and positive semidefinite, the  $d \times d'$  matrix  $\sigma$  such that  $a = \sigma \sigma^T$  exists and we thus find that a standard Brownian motion on a Riemannian manifold is in local coordinates given by the solution of the stochastic differential equation

$$dX_t^i = -\frac{1}{2} g^{jk}(X_t) \Gamma_{jk}^i(X_t) dt + \sigma_{ij}(X_t) dW_t^j, \quad i = 1, \dots, d \quad (3.4)$$

We aim to find the transition density of this process, as this leads us to a likelihood for the starting point as seen in equation (2.38). To achieve this, we require that

$$|b(t, x)| + |\sigma(t, x)| \leq C(1 + |x|) \quad (3.5)$$

$$|b(t, x) - b(t, y)| + |\sigma(t, x) - \sigma(t, y)| \leq D|x - y| \quad (3.6)$$

Now let us translate this to local coordinates using equation (3.3).

$$\left| \frac{1}{2} g^{jk}(x) \Gamma_{jk}^i(x) \right| + |\sigma(x)| \leq C(1 + |x|) \quad (3.7)$$

$$\left| \frac{1}{2} g^{jk}(x) \Gamma_{jk}^i(x) - \frac{1}{2} g^{jk}(y) \Gamma_{jk}^i(y) \right| + |\sigma(x) - \sigma(y)| \leq D|x - y| \quad (3.8)$$

Now notice that the functions  $g_{ij}$  are smooth functions on  $\mathcal{M}$  by definition and thus, as  $(gg^{-1})_{ij} = g_{ik}g^{kj} = \mathbb{1}_{\{i=j\}}$ ,  $g^{ij}$  must be smooth functions as well. Since the Christoffel symbols are a linear combination of the derivatives of  $g$ , these are smooth as well and thus the Lipschitz conditions are satisfied.

**Example 3.1.1** (The circle  $\mathbb{S}^1$ ). We know from example 2.1.42 that under the parameterization  $u \mapsto (\cos u, \sin u)$  we have that  $\Delta_{\mathbb{S}^1} = \frac{\partial^2}{\partial u^2}$ . When comparing this to equation (3.2), we quickly find that  $\sigma = a = I$ . We therefore easily find that a Brownian motion on a circle is parameterized by a process for the angle  $U_t$  that solves the stochastic differential equation  $dU_t = dW_t$  and thus, assuming  $U_0 = u_0$ , the position on the circle at time  $t$  is given by  $(X_t, Y_t) = (\cos(u_0 + W_t), \sin(u_0 + W_t))$ .

**Example 3.1.2** (The sphere  $\mathbb{S}^2$ ). The Laplace-Beltrami operator on the unit sphere, was found in equation (2.22) under the parameterization  $(u, v) \mapsto (\cos u \sin v, \sin u \sin v, \cos v)$  to be

$$(\Delta_{\mathbb{S}^2} f)(u, v) = \frac{1}{\sin^2 v} \frac{\partial^2 f}{\partial u^2} + \frac{\partial^2 f}{\partial v^2} + \frac{1}{\tan v} \frac{\partial f}{\partial v}.$$

Hence

$$(a_{ij}) = \begin{pmatrix} \frac{1}{\sin^2 v} & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 0 \\ \frac{1}{2 \tan v} \end{pmatrix}.$$

We find that a Brownian Motion on a sphere is parameterized by angles satisfying the following stochastic differential equation

$$d \begin{pmatrix} U_t \\ V_t \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{1}{2 \tan V_t} \end{pmatrix} dt + \begin{pmatrix} \frac{1}{|\sin V_t|} & 0 \\ 0 & 1 \end{pmatrix} dW_t \quad (3.9)$$

**Example 3.1.3** (The torus  $\mathbb{T}^2$ ). Consider the 2-torus  $\mathbb{T}^2 \subseteq \mathbb{R}^3$ . The torus  $\mathbb{T}^2$  is isomorphic to  $\mathbb{S}^1 \times \mathbb{S}^1$  and can be parameterized by

$$x(u, v) = ((R + r \cos v) \cos u, (R + r \cos v) \sin u, r \sin v) \quad (3.10)$$

where  $u, v \in [0, 2\pi]$  and  $R > r$ . Similarly as to the circle and the sphere, we can inherit a Riemannian metric on  $\mathbb{T}^2$  from  $\mathbb{R}^3$  that is given by

$$(g_{ij}) = \begin{pmatrix} \frac{\partial x}{\partial u} \cdot \frac{\partial x}{\partial u} & \frac{\partial x}{\partial u} \cdot \frac{\partial x}{\partial v} \\ \frac{\partial x}{\partial v} \cdot \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \cdot \frac{\partial x}{\partial v} \end{pmatrix} = \begin{pmatrix} (R + r \cos v)^2 & 0 \\ 0 & r^2 \end{pmatrix} \quad (3.11)$$

Standard calculations now show that the Christoffel symbols for the Levi-Civita connection are given by

$$\begin{aligned} \Gamma_{11}^1 &= 0 & \Gamma_{11}^2 &= \frac{(R + r \cos v) \sin v}{r} \\ \Gamma_{12}^1 &= -\frac{r \sin v}{R + r \cos v} & \Gamma_{12}^2 &= 0 \\ \Gamma_{21}^1 &= -\frac{r \sin v}{R + r \cos v} & \Gamma_{21}^2 &= 0 \\ \Gamma_{22}^1 &= 0 & \Gamma_{22}^2 &= 0 \end{aligned} \quad (3.12)$$

By applying equation (3.3), we quickly obtain the SDE describing a Brownian motion on  $\mathbb{T}^2$ .

$$d \begin{pmatrix} U_t \\ V_t \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{\sin V_t}{2r(R+r \cos V_t)} \end{pmatrix} dt + \begin{pmatrix} \frac{1}{|R+r \cos V_t|} & 0 \\ 0 & \frac{1}{r} \end{pmatrix} dW_t \quad (3.13)$$

where  $(W_t)_t$  is a standard  $\mathbb{R}^2$ -valued Brownian motion.

### 3.1.2. Discussion of this method

The characterization of Brownian motion via local coordinates discussed in this section has the advantage that it can be derived on any Riemannian manifold, as the only thing that we need are the Laplace-Beltrami operator and the Levi-Civita connection. In our examples we demonstrated the method with manifolds that are embedded in  $\mathbb{R}^3$  but this is not even a necessary condition. A disadvantage does show up on closer inspection of the SDEs that represent the Brownian motions.

Let us stick to the example of the sphere. In equation (3.9), we notice that a singularity appears when  $V_t$  approaches 0 or  $\pi$ , the coordinates that correspond to the north and south pole. When the Brownian motion approaches the north or south pole, the drift approaches  $-\infty$  or  $\infty$  respectively, driving the Brownian motion back in the other direction. To compensate for that, the diffusion coefficient of the other angle  $U_t$  approaches  $\infty$  as well. The Brownian motion is thus forced to stay on the sphere, so we do not immediately notice this behavior in simulations. It is however problematic, as it suggests that the assumptions 2.3.1 and 2.3.3 are not met, implying that a transition density might not exist.

The reason that these singularities appear is evident, the sphere is parameterized via one chart that does not include these points, while we should have two charts. Switching between charts can be tricky when simulating solutions to SDEs on manifolds, as the process should be continuous. In the next section we will therefore discuss another characterization of Brownian motion on manifolds that does not require local coordinates and avoids this problem.

## 3.2. Regular Submanifolds of $\mathbb{R}^N$ and Projections

In this section, we consider another characterization of Brownian motion on manifolds that is coordinate-free. As we briefly discussed in the previous section, and as we demonstrate later on in this thesis, this will help in simulations as it turns out that problems arise when we have to switch charts. In this chapter we derive such a method based on Rogers and Williams, 2000 [20], which generalizes the work of van den Berg and Lewis, 1985 [27], where similar results are derived on the sphere.

### 3.2.1. Regular submanifolds of $\mathbb{R}^N$

Note that we have considered manifolds, such as the sphere, as a  $d$ -dimensional space so far. For the sphere however, we can also use properties of  $\mathbb{R}^3$ , as this manifold is a subspace of  $\mathbb{R}^3$ . This section only focuses on manifolds that are subspace of  $\mathbb{R}^N$  of dimension  $d$ , where we set  $N = d + n > d$ . We start with the formal definition.

**Definition 3.2.1** (Regular submanifolds). *A set  $\mathcal{M} \subseteq \mathbb{R}^N$  is a regular  $\mathcal{C}^\infty$ -submanifold of  $\mathbb{R}^N$  of dimension  $d$  if for each  $x \in \mathcal{M}$ , there exists an open neighborhood  $G \subseteq \mathcal{M}$  of  $x$  and a  $\mathcal{C}^\infty$  function  $F : \mathbb{R}^d \rightarrow \mathbb{R}^n$  such that*

$$G = \left\{ \begin{pmatrix} y \\ F(y) \end{pmatrix} : y \in A \right\}$$

for an open set  $A \subseteq \mathbb{R}^d$ .

Notice that a regular submanifold of  $\mathbb{R}^N$  is indeed a  $d$ -dimensional manifold in the sense of definition 2.1.1, since we can take any point  $x = \begin{pmatrix} y \\ F(y) \end{pmatrix} \in \mathcal{M}$  and define the chart as the projection map  $G \rightarrow A$  that maps  $x$  to  $y$ .

In order to see that we can write many well-known manifolds in this form, we continue with an example.

**Example 3.2.2.** *Suppose that  $f : \mathbb{R}^N \rightarrow \mathbb{R}$  is a smooth function and that  $\nabla f$  does not vanish on  $\mathcal{M} := f^{-1}(\{0\})$ . Then it is a direct consequence of the implicit function theorem (see proposition 5.15*

of Lee, 2013 [13]) that  $\mathcal{M}$  is a regular submanifold of  $\mathbb{R}^N$  of dimension  $N - 1$ . Generally, a function  $f : \mathbb{R}^N \rightarrow \mathbb{R}^n$  that has a Jacobian matrix that is of full rank on  $\mathcal{M} = f^{-1}(\{0\})$  defines a regular submanifold of dimension  $N - n$ .

Relevant examples of such regular submanifolds are the  $d$ -spheres given by

$$\mathbb{S}^d = \{x \in \mathbb{R}^{d+1} : \|x\|^2 - 1 = 0\}$$

For now, let us make the following assumption

**Assumption 3.2.3.** *The manifold  $\mathcal{M}$  is of the form*

$$\mathcal{M} = \left\{ \begin{pmatrix} y \\ F(y) \end{pmatrix} : y \in \mathbb{R}^d \right\}$$

Since we have that any regular submanifold is locally of this form, we will be able to extend this assumption to general regular submanifolds.

Let  $J(y)$  denote the Jacobian matrix of  $F$  at  $y \in \mathbb{R}^d$  of dimension  $n \times d$  with entries given by

$$(J(y))_{i,j} = \frac{\partial F^i}{\partial y^j} \quad \text{for } i = 1, \dots, n \text{ and } j = 1, \dots, d.$$

Now suppose we have a curve in  $\mathcal{M}$ , given by  $t \mapsto x(t) = \begin{pmatrix} y(t) \\ F(y(t)) \end{pmatrix}$ . We can find a tangent vector  $v(t) = y'(t)$  to the curve  $t \mapsto y(t)$  in  $\mathbb{R}^d$  and then lift that vector via the chain rule to the tangent vector  $u(t)$  to  $x(t)$ .

$$u(t) := x'(t) = \begin{pmatrix} I \\ J(y(t)) \end{pmatrix} v(t)$$

that is tangent to  $\mathcal{M}$  at  $x(t)$ . We now deduce that when fixing  $y \in \mathbb{R}^d$  and  $x = \begin{pmatrix} y \\ F(y) \end{pmatrix} \in \mathcal{M}$ , we can expand a tangent vector  $v \in T_y \mathbb{R}^d$  to

$$u := \begin{pmatrix} I \\ J(y) \end{pmatrix} v = \begin{pmatrix} v^1 \\ \vdots \\ v^d \\ \sum \frac{\partial F^1}{\partial y^i} v^i \\ \vdots \\ \sum \frac{\partial F^n}{\partial y^i} v^i \end{pmatrix} \in T_x \mathcal{M}. \quad (3.14)$$

Since our manifold is of dimension  $d$ , we will now equip  $\mathbb{R}^d$  with a Riemannian metric that preserves the norms when projecting from  $\mathbb{R}^N$  to  $\mathbb{R}^d$ , since, when we accomplish that, the representations of vectors in  $\mathbb{R}^N$  behave the same as their representations in  $\mathbb{R}^d$ . We thus require that

$$\|u\|^2 = u^T u = v^T g(y) v. \quad (3.15)$$

We now observe that equation (3.15) is satisfied when

$$g(y) = \begin{pmatrix} I & J^T(y) \end{pmatrix} \begin{pmatrix} I \\ J(y) \end{pmatrix} = I + J(y)^T J(y) \quad (3.16)$$

### 3.2.2. Projections and Brownian motion

As the name of this section suggests, we use projections to characterize Brownian motion. More specifically, we use projections of  $\mathbb{R}^N$  onto tangent planes of the manifolds. Now that we have derived a Riemannian metric on these manifolds, we can further specify these projections.

**Theorem 3.2.4.** Let  $\pi$  denote the canonical projection  $\mathbb{R}^N \rightarrow \mathbb{R}^d$  given by  $\pi(x^1, \dots, x^N) = (x^1, \dots, x^d)$  and set  $y = \pi(x) \in \mathbb{R}^d$ . An orthogonal projection  $P(x)$  of  $\mathbb{R}^N$  onto  $T_x\mathcal{M}$  is given by

$$P(x) = \begin{pmatrix} I \\ J(y) \end{pmatrix} g(y)^{-1} \begin{pmatrix} I & J(y)^T \end{pmatrix} \quad (3.17)$$

*Proof.* By definition,  $g(y) \in \mathbb{R}^{d \times d}$ , which implies that for any  $\xi \in \mathbb{R}^N$ ,  $g(y)^{-1} \begin{pmatrix} I & J(y)^T \end{pmatrix} \xi \in \mathbb{R}^d$  and it thus follows from equation (3.14) that  $P(x)$  indeed maps to  $T_x\mathcal{M}$ . When substituting equation (3.16) into equation (3.17), one can also quickly see that  $P(x)^2 = P(x)$  and thus  $P(x)$  is indeed an orthogonal projection onto  $T_x\mathcal{M}$ .  $\square$

**Corollary 3.2.5.** Let  $f : \mathbb{R}^N \rightarrow \mathbb{R}$  be a smooth function that does not vanish on  $\mathcal{M} := f^{-1}(\{0\})$ . Then  $\mathcal{M}$  is a regular submanifold of  $\mathbb{R}^N$  and the projection  $P(x)$  onto  $T_x\mathcal{M}$  is given by

$$P(x) = I - n(x)n(x)^T \quad (3.18)$$

where  $n(x) = \frac{\nabla f(x)}{|\nabla f(x)|}$  denotes the normal vector to  $T_x\mathcal{M}$ .

*Proof.* This corollary can be directly deduced from theorem 3.2.4, but we instead give a more intuitive proof. Since  $\mathbb{R}^N$  is a Euclidean space and both  $P(x)$  and  $I - n(x)n(x)^T$  are linear maps, we only need to prove the result for a basis of  $\mathbb{R}^N$ . Now let  $\xi_1, \dots, \xi_{N-1}$  denote an orthonormal basis for  $T_x\mathcal{M}$ . Then  $\xi_1, \dots, \xi_{N-1}, n(x)$  is an orthonormal basis of  $\mathbb{R}^N$ . Since  $n(x)^T n(x) = 1$  and  $n(x)^T \xi_i = 0$  for all  $i$ , we quickly deduce that  $\begin{pmatrix} I - n(x)n(x)^T \end{pmatrix} \xi_i = \xi_i$  and  $\begin{pmatrix} I - n(x)n(x)^T \end{pmatrix} n(x) = 0$  and thus  $I - n(x)n(x)^T$  is indeed an orthogonal projection onto  $T_x\mathcal{M}$ .  $\square$

We can now use the projection of  $\mathbb{R}^N$  onto tangent spaces to construct a Brownian motion on regular submanifolds via the following theorem.

**Theorem 3.2.6.** Let  $\mathcal{M}$  be a regular submanifold of  $\mathbb{R}^N$  of the form of assumption 3.2.3, let  $W$  be a Brownian motion in  $\mathbb{R}^N$  and let  $X$  be the solution of Stratonovich stochastic differential equation given by

$$dX_t = P(X_t) \circ dW_t, \quad X_0 \in \mathcal{M}. \quad (3.19)$$

Then  $X$  takes values in  $\mathcal{M}$ . Furthermore, the infinitesimal generator  $\mathcal{L}$  of the process  $\pi(X)$  in the Riemannian manifold  $(\mathbb{R}^d, g)$  is given by  $\mathcal{L} = \frac{1}{2}\Delta$ , where  $\Delta$  denotes the Laplace-Beltrami operator.

*Proof.* Throughout this proof, we let  $Y$  denote the  $\mathbb{R}^d$ -valued process given by  $Y_t = \pi(X_t)$ . We first need to verify that  $X$  indeed takes values in  $\mathcal{M}$ . We thus need to verify that  $\tilde{X}_t := (X_t^{d+1}, \dots, X_t^N) = F(Y_t)$ . Now recall from proposition 2.2.15 that Stratonovich SDEs follow the classical rules of calculus, so we have that  $dF(Y_t) = J(Y_t) \circ dY_t$ . We can now see that for  $1 \leq r \leq n$

$$dF^r(Y_t) = J(Y_t)_{r,i} \circ dY_t^i$$

where we used Einsteins summation convention to sum  $i$  from 1 up to  $d$ .

Now note that  $Y$  denotes the first  $d$  rows of the process  $X$  that solves equation (3.19) and that it thus follows from equation (3.17) when substituting  $P(X_t)$  that

$$dY_t = g(Y_t)^{-1} \begin{pmatrix} I & J(Y_t)^T \end{pmatrix} \circ dW_t,$$

We thus find that

$$dF(Y_t) = J(Y_t) \circ dY_t = J(Y_t) g(Y_t)^{-1} \begin{pmatrix} I & J(Y_t)^T \end{pmatrix} \circ dW_t \quad (3.20)$$

Similarly, note that  $\tilde{X}$  denotes the last  $n$  rows of the process  $X$  and that it thus also follows from equation (3.17) when substituting  $P(X_t)$  that

$$d\tilde{X}_t = J(Y_t) g(Y_t)^{-1} \begin{pmatrix} I & J(Y_t)^T \end{pmatrix} \circ dW_t \quad (3.21)$$

Combining the last two equations (3.20) and (3.21) now gives us that

$$d(X_t^{d+r} - F^r(Y_t)) = 0, \quad \text{for } r \in \{1, \dots, n\}.$$

Since  $X_0 \in \mathcal{M}$ , we have that  $X_0 = \begin{pmatrix} Y_0 \\ F(Y_0) \end{pmatrix}$  and thus we can conclude that

$$\mathbb{P}\left(X_t = \begin{pmatrix} Y_t \\ F(Y_t) \end{pmatrix} \text{ for all } t\right) = 1.$$

The process  $X$  thus stays on  $\mathcal{M}$  almost surely and we have yet to prove that the infinitesimal generator of  $Y$  is given by  $\frac{1}{2}\Delta$ . Since  $dY_t = g(Y_t)^{-1}(I - J(Y_t)^T) \circ dW_t$ , we may write

$$dY_t^i = \sigma_{ij}(Y_t) \circ dW_t^j, \quad \text{where } \sigma_{ij}(y) = \begin{cases} g^{ij}(y) & \text{for } 1 \leq j \leq d \\ g^{ik}(y) J_{jk}(y) & \text{for } d+1 \leq j \leq N \end{cases}. \quad (3.22)$$

In order to find the infinitesimal generator, we write equation (3.22) in Itô form.

$$dY_t^i = \frac{1}{2} \sum_{j,k} \sigma_{kj}(Y_t) \partial_k \sigma_{ij}(Y_t) dt + \sigma_{ij}(Y_t) dW_t^j \quad (3.23)$$

From equation (3.16), it follows that  $\sigma\sigma^T = g^{-1}(I + J^T J)g^{-1} = g^{-1}$ . We now recall from 2.2.24 that this now gives us that the infinitesimal generator  $\mathcal{L}$  of the process  $Y$  can be written as

$$\mathcal{L} = \frac{1}{2} g^{ij} \partial_i \partial_j + b^i \partial_i \quad (3.24)$$

where

$$b^i = \frac{1}{2} \sum_{j \leq d, k} \sigma_{kj} \partial_k \sigma_{ij} + \frac{1}{2} \sum_{j > d, k} \sigma_{kj} \partial_k \sigma_{ij}. \quad (3.25)$$

Substituting equation (3.22), results in

$$2b^i = \sum_{j \leq d, k} g^{kj} \partial_k g^{ij} + \sum_{j > d, k} g^{k\ell} J_{j\ell} \partial_k (g^{i\ell} J_{j\ell})$$

Now notice that the Riemannian metric  $g$  as given in equation (3.16) immediately gives us the identity

$$g^{-1} + g^{-1} J^T J = I.$$

Applying the operator  $\partial_k$  to both sides and applying the product rule properly leads to

$$\partial_k g^{-1} + \partial_k (g^{-1} J^T) J = - (g^{-1} J^T) \partial_k J. \quad (3.26)$$

An important observation is that if we fill in  $g = I + J^T J$  into the definition of the Christoffel Symbols and use that  $\partial_k J_{j\ell} = \partial_\ell J_{jk}$ , we obtain that

$$\begin{aligned} \Gamma_{jk}^i &= \frac{1}{2} g^{i\ell} (\partial_j g_{k\ell} + \partial_k g_{\ell j} - \partial_\ell g_{jk}) \\ &= \frac{1}{2} g^{i\ell} \sum_r \{ \partial_j (J_{rk} J_{r\ell}) + \partial_k (J_{r\ell} J_{rj}) - \partial_\ell (J_{rj} J_{rk}) \} \\ &= \frac{1}{2} g^{i\ell} \sum_r \{ J_{rk} \partial_j J_{r\ell} + (\partial_j J_{rk}) J_{r\ell} + J_{r\ell} \partial_k J_{rj} + (\partial_k J_{r\ell}) J_{rj} - J_{rj} \partial_\ell J_{rk} - (\partial_\ell J_{rj}) J_{rk} \} \\ &= g^{i\ell} \sum_r J_{r\ell} \partial_k J_{rj} \end{aligned}$$

Hence, if we wish to evaluate equation (3.26) in a single pair  $(i, j)$ , we obtain the formula

$$\partial_k g^{ij} + \sum_\ell \partial_k (g^{im} J_{\ell m}) J_{\ell j} = - \sum_r g^{i\ell} J_{r\ell} \partial_k J_{rj} = -\Gamma_{jk}^i.$$

Combining this result with equation (3.25) results in

$$2b^i = -g^{jk}\Gamma_{jk}^i.$$

and we thus have that

$$\mathcal{L} = \frac{1}{2} \left( g^{ij} \partial_i \partial_j - g^{jk} \Gamma_{jk}^i \partial_i \cdot \right)$$

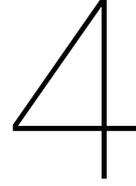
After revisiting equation (2.21), we now conclude that the process  $Y$  is generated by the operator  $\mathcal{L} = \frac{1}{2} \Delta$ .  $\square$

### 3.2.3. Discussion of this method

At this point we have constructed a way to characterize Brownian motion on regular submanifolds. One can think of these manifolds as vector spaces in  $\mathbb{R}^N$  that have  $d$  "free" coordinates and  $n$  constraints that are represented in the function  $F$ . If we equip  $\mathbb{R}^d$  with a Riemannian metric that the "free" coordinates behave the same in  $\mathbb{R}^d$  as the points in  $\mathbb{R}^N$  do on the manifold, we can construct a stochastic process  $Y$  that has the Laplace-Beltrami operator on  $\mathbb{R}^d$  belonging to that Riemannian metric as infinitesimal generator and, when using the constraints in  $F$  to project this process to  $\mathbb{R}^N$  via  $y \mapsto (y, F(y))$ , generates Brownian motion on  $\mathcal{M}$ .

Previously, we have derived a method via local coordinates by characterizing Brownian motion on  $\mathcal{M}$  as the process that is generated by the Laplace-Beltrami operator in local coordinates. A noticeable difference with this method is that, when the projection matrix onto a tangent plane is known in closed form, local coordinates will not be needed anymore, which justifies the notion of calling this a "coordinate-free" characterization. As we saw in corollary 3.2.5, there are many examples where we know the projection matrix in closed form. This is an advantage of this method compared to the generation of Brownian motion via the Laplace-Beltrami operator. There are also disadvantages however, as this method requires  $N$  driving Brownian motions in  $\mathbb{R}^N$  in the SDE for Brownian motion on the manifolds, whereas the other method requires  $d' \leq d$ . Another disadvantage of this method is that it is only applicable on manifolds that are regular submanifolds of  $\mathbb{R}^N$ , so this method does not provide us with a framework for general Riemannian manifolds.





# Stochastic Simulation of Diffusions and diffusion bridges

In this chapter, some algorithms for the simulations of diffusion and diffusion bridges in  $\mathbb{R}^d$  are discussed.

## 4.1. The Euler-Maruyama method

Let  $X$  be a diffusion process written in integral form as

$$X_t = X_0 + \int_0^t b(s, X_s) ds + \int_0^t \sigma(s, X_s) dW_s$$

The most basic numerical scheme to approximate a path of this integral equation consists of discretizing time with equal time steps. Let therefore  $\Delta t > 0$  be the step size. Then we can use the forward Euler method to approximate the first integral as

$$\int_t^{t+\Delta t} b(s, X_s) ds \approx \Delta t b(t, X_t)$$

Now to find an approximation for the stochastic integral, it is important to notice that, by definition,  $\Delta W_t = W_{t+\Delta t} - W_t \sim \mathcal{N}(0, \Delta t)$ . A natural approximation is therefore given by

$$\int_t^{t+\Delta t} \sigma(s, X_s) dW_s \approx \sigma(t, X_t) \sqrt{\Delta t} Z$$

with  $Z \sim \mathcal{N}(0, 1)$ . Given  $X_t$ , we can thus approximate  $X_{t+\Delta t}$  by

$$X_{t+\Delta t} \approx X_t + \Delta t b(t, X_t) + \sigma(t, X_t) \sqrt{\Delta t} Z \quad (4.1)$$

Knowing  $X_0$ , we can now simulate the diffusion process on an arbitrary grid via the Euler-Maruyama method and perform a linear interpolation between the grid points. This method naturally extends to approximations for diffusions in higher dimensions by using multivariate normally distributed random variables instead of the standard normal  $Z$ . An extensive proof of strong convergence of this method when  $\Delta t \downarrow 0$  for space-dependent diffusions under the satisfying Lipschitz conditions (2.28) and (2.29) can be seen in theorem 2.2 in Higham et al, 2002 [8].

**Example 4.1.1** (Simulating standard Brownian motion). *A path of a standard Brownian motion can be simulated on a grid  $\{0 = t_0 < \dots < t_k = T\}$  by simulating standard normal random variables  $N_1, \dots, N_k$ , setting  $w_0 = 0$ ,  $w_{t_i} = w_{t_{i-1}} + \sqrt{t_i - t_{i-1}} N_i$  for  $i = 1, \dots, k$  and interpolating linearly between these points. Examples in  $\mathbb{R}$ ,  $\mathbb{R}^2$  and  $\mathbb{R}^3$  can be found in figures 2.2 and 2.3.*

**Example 4.1.2** (Simulating Brownian motion on the circle  $\mathbb{S}^1$ ). Recall the circle is one of the rare examples in which we know the exact solution of the SDE describing Brownian motion in local coordinates (see example 3.1.1). When we let the Brownian motion start in  $(1, 0)$ , this solution is given by

$$(X_t, Y_t) = (\cos(W_t), \sin(W_t)) \quad (4.2)$$

where  $W = (W_t)_t$  is a standard Brownian motion in  $\mathbb{R}$ . We can thus simulate Brownian motion on the circle by simulating a path of a standard Brownian motion as in example 4.1.1 and plugging it into equation (4.2).

Since we typically do not have a solution in closed form, it is interesting to see how the Euler-Maruyama method behaves when this solution is written as a 2D-SDE. By the Itô formula 2.30 to equation (4.2), we find that

$$d \begin{pmatrix} X_t \\ Y_t \end{pmatrix} = -\frac{1}{2} \begin{pmatrix} \cos W_t \\ \sin W_t \end{pmatrix} dt + \begin{pmatrix} -\sin W_t \\ \cos W_t \end{pmatrix} dW_t = -\frac{1}{2} \begin{pmatrix} X_t \\ Y_t \end{pmatrix} dt + \begin{pmatrix} -Y_t \\ X_t \end{pmatrix} dW_t \quad (4.3)$$

Note that we still consider the driving Brownian motion  $W$  to be in  $\mathbb{R}$ . We can now apply the Euler-Maruyama method to simulate a path of the solution to this equation with starting point  $(1, 0)$ . The simulated paths of  $X_t$  and  $Y_t$  can be found in figures 4.1 and 4.2.

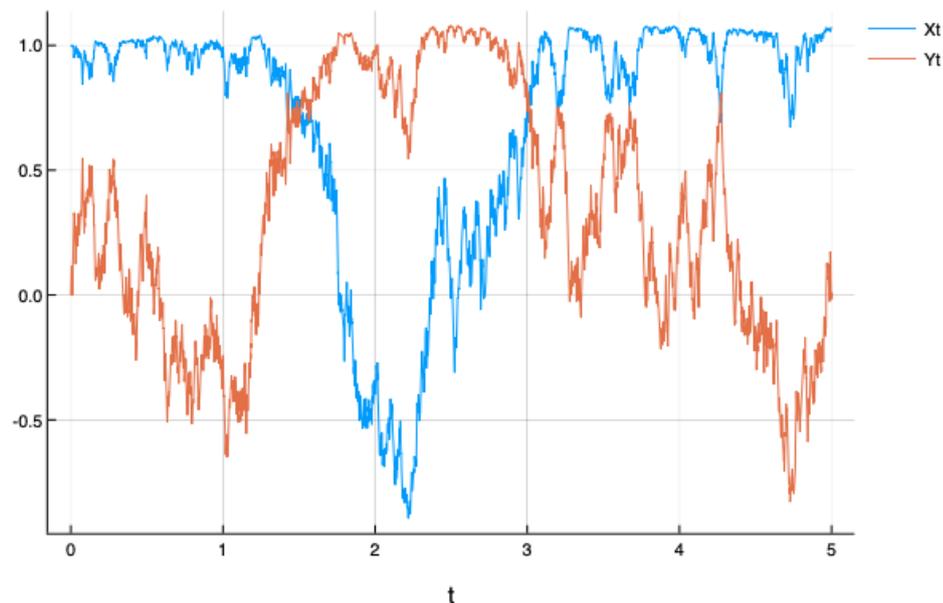


Figure 4.1: The  $x$ - and  $y$ -coordinates for a Brownian motion on a circle simulated via the Euler-Maruyama method with a time step of  $\frac{1}{1000}$ .

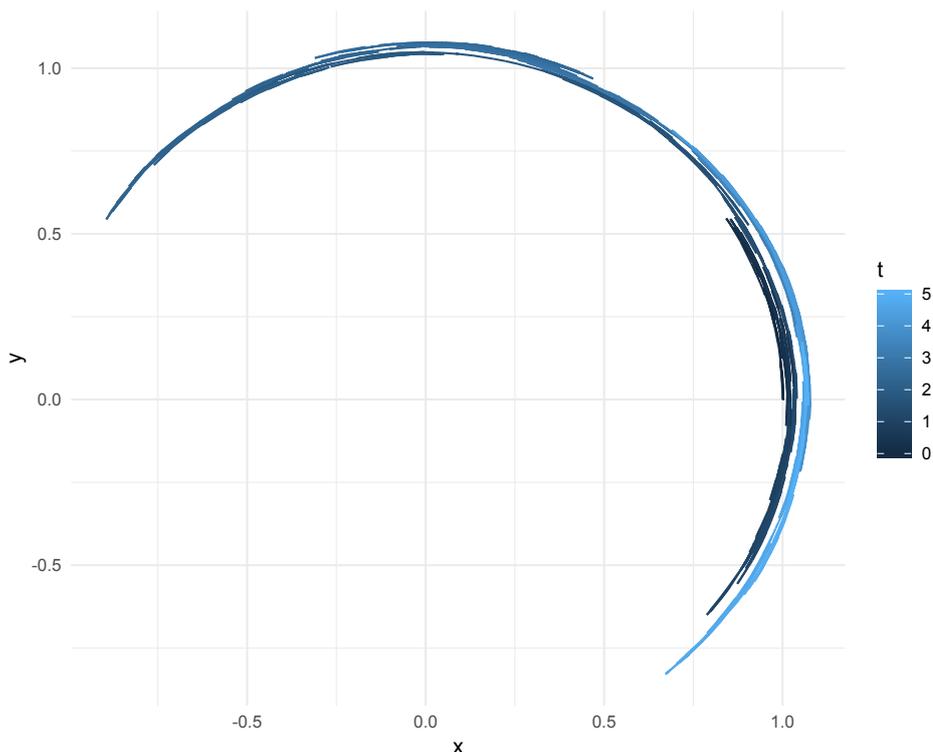


Figure 4.2: The simulated Brownian motion on the circle corresponding to figure 4.1 shown in  $\mathbb{R}^2$

As can be seen in figure 4.2, the simulated Brownian motion does not stay on the circle as it should. The reason that the Brownian motion does not stay on the circle has to do with the numerical approximation. The law of the process we simulate is close to the law of the actual Brownian motion on the circle, but not equal as we do discretize time. Typically, lower values of  $\Delta t$  lead to better approximations but result in longer computation times.

**Example 4.1.3.** On the sphere, we can apply this method as well and simulate the SDE given in equation (3.9). By again applying the Itô formula, we can see how the Euler-Maruyama method performs on a three-dimensional system of SDEs as well.

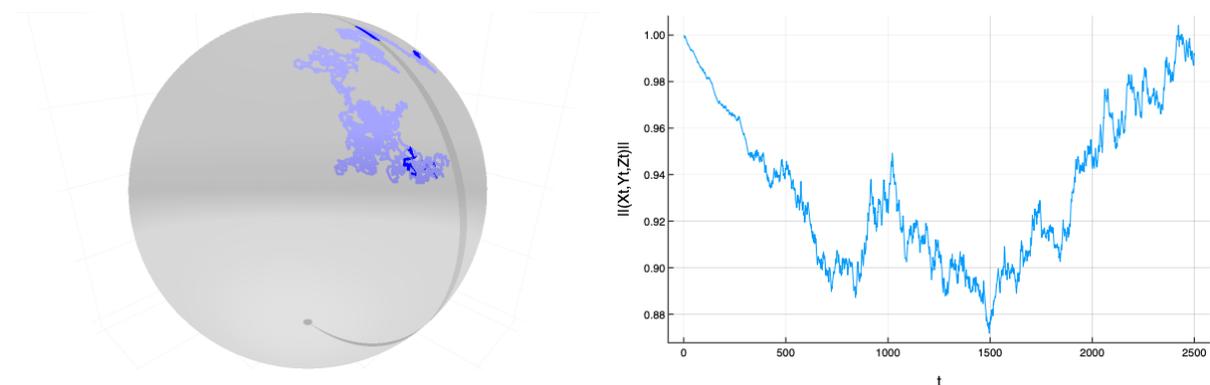


Figure 4.3: A Brownian motion on a sphere, simulated using the Euler-Maruyama method On the left, the path is plotted on a sphere and on the right, the norm of the simulated path is demonstrated.

On the sphere we see the same behavior as we saw on the circle.

**Example 4.1.4** (Brownian motion on the 2-torus  $\mathbb{T}^2$ ). We can also apply Euler-Maruyama on equation (3.13) to get a visualization of a Brownian motion on the torus as seen in figure 4.4.

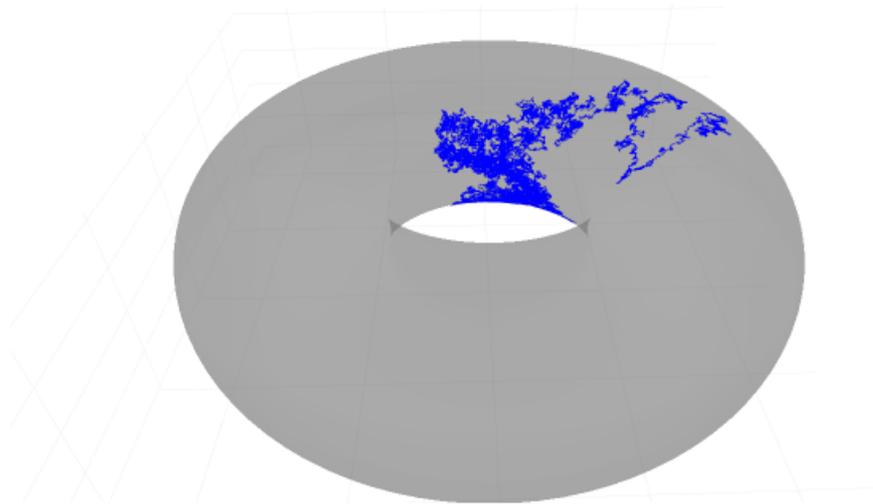


Figure 4.4: A simulated path of Brownian motion on the 2-torus, simulated using the Euler-Maruyama method.

## 4.2. Using 2-jets and the exponential map

The second method we propose for simulating Brownian motion is seen in Armstrong and Brigo, 2018 [1]. In this section we briefly introduce their results and use them for the simulation of Brownian motion. Here, it is supposed that each point  $x \in \mathbb{R}^d$  has an associated smooth curve  $\gamma_x : \mathbb{R} \rightarrow \mathbb{R}^d$  with  $\gamma_x(0) = x$ . We can now define a stochastic relation by choosing a time step  $\Delta t > 0$  and setting

$$X_0^{\Delta t} = x_0, \quad X_{t+\Delta t}^{\Delta t} = \gamma_{X_t^{\Delta t}}(W_{t+\Delta t} - W_t) \quad (4.4)$$

where  $(W_t)_t$  is a standard  $\mathbb{R}$ -valued Brownian motion. Note that since  $W_{t+\Delta t} - W_t \sim \mathcal{N}(0, \Delta t)$ , we can derive a stochastic process by defining the trajectory between  $X_t^{\Delta t}$  and  $X_{t+\Delta t}^{\Delta t}$  via the curve  $s \mapsto \gamma_{X_t^{\Delta t}}(s)$  where  $s$  ranges from  $s = 0$  to  $s = \epsilon_t \sqrt{\Delta t}$  where  $\epsilon_t \sim \mathcal{N}(0, 1)$ .

As  $\gamma_x$  is a smooth curve, we can consider a Taylor expansion at 0, that is given by

$$\gamma_x(t) = x + \gamma'_x(0)t + \frac{1}{2}\gamma''_x(0)t^2 + R_x t^3, \quad \text{where } R_x = \frac{1}{6}\gamma'''_x(\xi) \text{ for some } \xi \in [0, t].$$

Substituting the Taylor expansion in equation (4.4) and setting  $\Delta X_t^{\Delta t} = X_{t+\Delta t}^{\Delta t} - X_t^{\Delta t}$  and  $\Delta W_t = W_{t+\Delta t} - W_t$ , results in

$$\Delta X_t^{\Delta t} = \gamma'_{X_t^{\Delta t}}(0) \Delta W_t + \frac{1}{2}\gamma''_{X_t^{\Delta t}}(0) (\Delta W_t)^2 + R_{X_t^{\Delta t}} (\Delta W_t)^3$$

Now, by replacing  $(\Delta W_t)^2$  by  $\Delta t$  (since the Itô formula states that  $dW_t dW_t = d[W]_t = dt$ ) and neglecting higher order terms when  $\Delta t$  gets small, we can expect that when  $\Delta t \rightarrow 0$ , we obtain convergence to the SDE given by

$$dX_t = b(X_t) dt + \sigma(X_t) dW_t$$

where we define  $b(X) = \frac{1}{2}\gamma''_X(0)$  and  $\sigma(X) = \gamma'_X(0)$ . We will specify this notion of convergence in theorem 4.2.2.

**Definition 4.2.1** (*k*-jets). A *k*-jet of a function between smooth manifolds  $\mathcal{M}$  and  $\mathcal{N}$  is the equivalence class of all smooth maps  $\mathcal{M} \rightarrow \mathcal{N}$  that are equal up to order *k* in one, and hence all, coordinate systems. Given a curve  $\gamma_x$ , we will denote the *k*-jet associated with  $\gamma_x$  by  $j_k(\gamma_x)$ .

We now go into the main result of the publication of Armstrong and Brigo stating the conditions that guarantee the convergence suggested previously. We state the results for  $\mathbb{R}^d$ -valued diffusions driven by  $\mathbb{R}^{d'}$ -valued Brownian motion. For the intermediate results, we refer to the publication [1].

**Theorem 4.2.2.** Let  $\gamma_x : \mathbb{R}^d \rightarrow \mathbb{R}^d$  be a smoothly varying family of functions with first and second derivatives in  $\mathbb{R}^{d'}$  satisfy that Lipschitz conditions around 0. Furthermore suppose that we have a uniform bound on all the third derivatives in 0. Now let  $T$  be a fixed time, let  $\mathcal{T}^N = \{0, \Delta t, 2\Delta t, \dots, N\Delta t = T\}$  and let  $X_t$  denote the 2-jet scheme defined by

$$X_{t+\epsilon} = \gamma_{X_t} \left( \frac{\epsilon}{\Delta t} (W_{t+\Delta t} - W_t) \right), \quad t \in \mathcal{T}^{N-1}, \epsilon \in [0, \Delta t], X_0 = x_0.$$

As  $\Delta t \downarrow 0$ , this process converges in  $L^2(\mathbb{P})$  to the process  $(\tilde{X}_t)_t$  given by

$$\tilde{X}_t = \tilde{X}_0 + \int_0^t b(\tilde{X}_s) ds + \sum_{\alpha=1}^{d'} \int_0^t \sigma_\alpha(\tilde{X}_s) dW_s^\alpha$$

where

$$b(x) = \frac{1}{2} \sum_{\alpha=1}^m \frac{\partial^2 \gamma_x}{\partial u^\alpha \partial u^\alpha} \Big|_{u=0} \quad \text{and} \quad \sigma_\alpha(x) = \frac{\partial \gamma_x}{\partial u^\alpha} \Big|_{u=0}$$

*Proof.* See Theorem 2.4 of Armstrong and Brigo, 2018 [1]. □

What this method proves is that we have a family of functions that combine a starting point and then tell us, based on a Brownian increment where our stochastic process should go. Given an SDE, we can thus find a curve that fits the drift and diffusion coefficient via its derivatives and simply fill in increments of Brownian motion.

When simulating Brownian motion on a Riemannian manifold, the most straightforward choice for  $\gamma_x$  is the exponential map, especially when a closed form of this map exists, as is the case on the sphere (see 2.1.35). Intuitively, this means that we plug in a starting value  $X_t$ , project the increment  $dW_t$  of a Brownian motion in the tangent plane  $T_{X_t}\mathcal{M}$  and transfer that to an increment on the manifold starting at  $X_t$  and moving in the direction of  $dW_t$ .

**Example 4.2.3** (The circle  $\mathbb{S}^1$ ). When again considering equation (4.3), we find that the 2-jet scheme given by

$$\gamma_{(x,y)}(t) = \left( x - yt - \frac{1}{2}xt^2, y + x - \frac{1}{2}yt^2 \right)$$

substituted in theorem 4.2.2 coincides with the SDE specified by equation (4.3). We can thus simulate a Brownian motion on the circle using the recursion  $X_{t+\Delta t} = \gamma_{X_t}(\Delta W_t)$ . Using the same time-step and the same Brownian increments as used when creating figure 4.2, we obtain the result found in figure 4.5.

Clearly using this method, the simulated Brownian motion seems to stay closer to the circle compared to the result from Euler-Maruyama method seen in figure 4.2.

The exponential map on the circle is given by  $\text{Exp}_p v = (\cos(\theta + v_0), \sin(\theta + v_0))$  where  $p = (\cos \theta, \sin \theta)$  and  $v = v_0(-\sin \theta, \cos \theta)$  (see example 2.1.33). The one dimensional Brownian motion can now be substituted in place of  $\theta$ . If we use this map to simulate Brownian motion on the circle, parameterized by  $X_t = (\cos \theta_t, \sin \theta_t)$ , we have that  $X_{t+\Delta t} = (\cos(\theta_t + \Delta W_t), \sin(\theta_t + \Delta W_t))$ . This means that the simulated Brownian motion on the circle comes down to simulating a standard  $\mathbb{R}$ -valued Brownian motion and plugging this into the actual solution  $X_t = (\cos(\theta_0 + W_t), \sin(\theta_0 + W_t))$ .

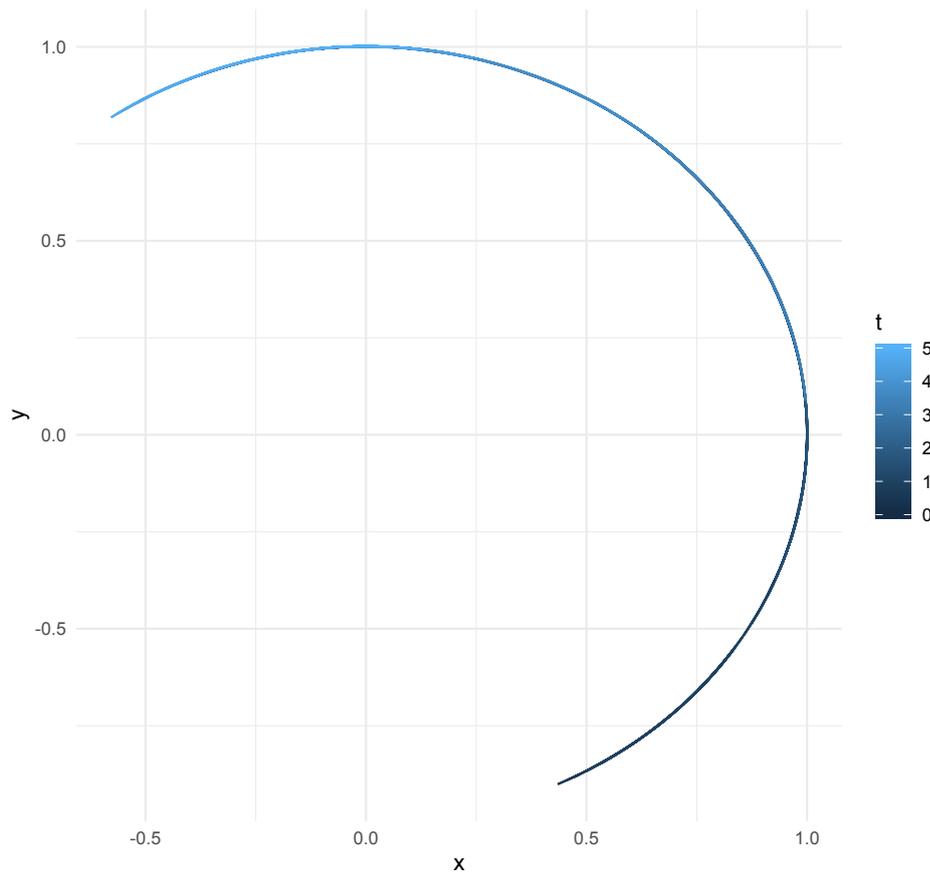


Figure 4.5: Brownian motion on the circle simulated using 2-jets.

**Example 4.2.4.** As we have seen in corollary 2.1.35, we have a closed form of the exponential map on the sphere as well. It is given by

$$\text{Exp}_p v = \begin{cases} p & \text{if } v = 0 \\ \cos(\|v\|) p + \sin(\|v\|) \frac{v}{\|v\|} & \text{if } \|v\| > 0 \end{cases}$$

This requires that the vector  $v \in T_p \mathbb{S}^2$ , which is generally not the case for an  $\mathbb{R}^2$ -valued Brownian motion. We can therefore take the vectors orthogonal to  $p$  given by  $e_1^p = (-p_2, p_1, 0)$  and  $e_2^p = p \times e_1^p$  and normalize them to find an orthonormal base for  $T_p \mathbb{S}^2$ . Given a Brownian increment in  $\mathbb{R}^2$ ,  $\Delta W = (\Delta W^1, \Delta W^2)$  resulting from increments of Brownian motions  $W^1$  and  $W^2$  in  $\mathbb{R}$ , a Brownian increment on  $T_p \mathbb{S}^2$  is now given by  $\Delta W^p = \Delta W^1 e_1^p + \Delta W^2 e_2^p$ . Next we can simulate forward using

$$X_{t+\Delta t} = \text{Exp}_{X_t}(\Delta W^{X_t}).$$

A sample path simulated using this method is given in figure 4.6.

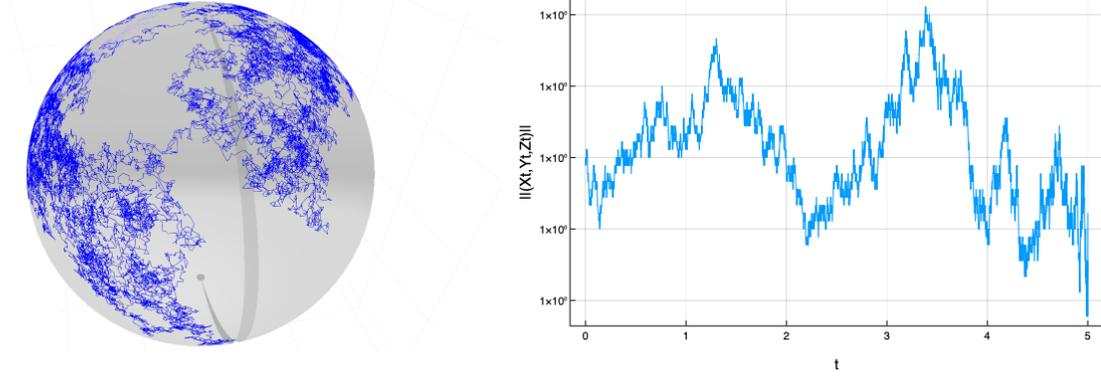


Figure 4.6: A sample path Brownian motion on a sphere simulated using the Exponential map (left) and its norm (right).

### 4.3. Simulation of diffusion bridges using guided proposals

A diffusion conditioned to hit a specified point at a future time is called a *diffusion bridge*. Various methods have been discussed for simulating diffusions and this section will continue on that by writing diffusion bridges as standard diffusions by altering the drift. In this section, we let  $X$  denote a time-inhomogeneous diffusion process that takes values in  $\mathbb{R}^d$  governed by the SDE

$$dX_t = b(t, X_t) dt + \sigma(t, X_t) dW_t, \quad X_0 = x. \quad (4.5)$$

Here  $b : [0, \infty) \times \mathbb{R}^d \rightarrow \mathbb{R}^d$  and  $\sigma : [0, \infty) \times \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d'}$  are the drift and diffusion coefficient and  $W$  is a  $d'$ -dimensional Brownian motion. Now let  $X^*$  denote the respective diffusion bridge, i.e. the conditioned stochastic process  $X^* := (X \mid X_T = v)$  conditioned to hit a given point  $v \in \mathbb{R}^d$  at a given time  $T > 0$ . Under weak assumptions (see Lyons and Zheng, 1990 [15]), we can represent the bridge  $X^*$  as the solution to the SDE

$$dX_t^* = b^*(t, X_t^*) dt + \sigma(t, X_t^*) dW_t, \quad X_0^* = x, \quad t \in [0, T) \quad (4.6)$$

where

$$b^*(t, x) = b(t, x) + a(t, x) \nabla_x \log p(t, x; T, v).$$

Here  $a$  denotes the matrix  $\sigma \sigma^T$  and  $p$  is the transition density of the process  $x$ , i.e.  $\mathbb{P}(X_t \in dy \mid X_s = x) = p(s, x; t, y) dy$  for  $s < t$ . The term  $a(t, x) \nabla_x \log p(t, x; T, v)$  is usually referred to as the guiding term. Since  $p$  is generally intractable, simulating the process seen in equation (4.6) is usually not possible and we thus need an approximation method.

There are multiple ways to approximate a solution to 4.6 and they usually involve a slight adjustment of the drift  $b^*$ . Delyon and Hu, 2006 [6] have considered proposals  $X^\nabla$  and  $X^\Delta$  that solve the SDEs given by

$$dX_t^\nabla = \left( b(t, X_t^\nabla) + \frac{v - X_t^\nabla}{T - t} \right) dt + \sigma(t, X_t^\nabla) dW_t$$

and

$$dX_t^\Delta = \frac{v - X_t^\Delta}{T - t} dt + \sigma(t, X_t^\Delta) dW_t$$

We choose to move to the guided proposals proposed in Schauer et al., 2017a [22], as they allow a bit more flexibility in altering the drift of equation (4.6). The guided proposal  $X^\circ$  is given by the solution of the SDE

$$dX_t^\circ = b^\circ(t, X_t^\circ) dt + \sigma(t, X_t^\circ) dW_t, \quad X_0^\circ = x, \quad t \in [0, T) \quad (4.7)$$

where

$$b^\circ(t, x) = b(t, x) + a(t, x) \nabla_x \log \tilde{p}(t, x; T, v).$$

and  $\tilde{p}$  is the transition density of a process of which it is known in closed form. In Schauer et al., 2017a [22], it is shown that we can choose  $\tilde{X}$  be a linear process, which is an example of a class of SDEs with tractable transition densities. A linear process  $\tilde{X}$  is governed by the SDE

$$d\tilde{X}_t = [\tilde{B}(t)\tilde{X}_t + \tilde{\beta}(t)] dt + \tilde{\sigma}(t) dW_t$$

We refer to  $\tilde{X}$  as the auxiliary process. It is shown by Schauer et al., 2017a [22], that when  $\tilde{\sigma}$  is such that  $\tilde{\sigma}(T) = \sigma(T, v)$  under some conditions on  $\tilde{B}$ ,  $\tilde{\beta}$  and  $\tilde{\sigma}$ , the drift of the original process  $X$  and the convergence of  $X^\circ$  to  $v$  as  $t \uparrow T$ , we have that the laws of  $X^*$  and  $X^\circ$  are equivalent on  $[0, T]$  and

$$\frac{d\mathbb{P}_T^*(X^\circ)}{d\mathbb{P}_T^\circ(X^\circ)} = \frac{\tilde{p}(0, x; T, v)}{p(0, x; T, v)} \psi(T, X^\circ). \quad (4.8)$$

Here  $\mathbb{P}_T$  represents the law of the process  $(X_t : t \in [0, T])$ . Before we are able to write down  $\psi$ , let us introduce additional notation.

$$R(s, x) = \log p(s, x), \quad r(s, x) = \nabla_x R(s, x), \quad H(s, x) = \left( -\frac{\partial^2 R}{\partial x^i \partial x^j}(s, x) \right)_{i,j} \quad (4.9)$$

where  $p(s, x) = p(s, x; T, v)$ . We similarly define  $\tilde{R}$ ,  $\tilde{r}$  and  $\tilde{H}$  for the auxiliary process  $\tilde{X}$ . Going back to equation (4.8), we have

$$\psi(t, X^\circ) = \exp \left( \int_0^t G(s, X_s^\circ) ds \right) \quad (4.10)$$

where, with  $\tilde{b}(t, x) = \tilde{B}(t)x + \tilde{\beta}(t)$ ,

$$G(s, x) = \left( b(s, x) - \tilde{b}(s, x) \right)^T \tilde{r}(s, x) - \frac{1}{2} \text{Trace} \left( [a(s, x) - \tilde{a}(s, x)] [\tilde{H}(s, x) - \tilde{r}(s, x)\tilde{r}(s, x)^T] \right)$$

It is important to notice that  $G$  does not depend on any unknown objects here. We can thus use the proposal  $X^\circ$  to simulate the diffusion bridge  $X^*$ . In practice one uses an Metropolis-Hastings (MH) algorithm (see appendix B). Note that, as functions of guided proposals  $X^\circ$ , we have  $\frac{d\mathbb{P}_T^*(X^\circ)}{d\mathbb{P}_T^\circ(X^\circ)} \propto \psi(T, X^\circ)$ . We can thus choose an acceptance rate  $\rho$  and use the following algorithm.

**Input:** Start with a standard Brownian motion  $W$  in  $\mathbb{R}^{d'}$  and simulate a solution  $X$  to equation (4.5) both simulated via the Euler-Maruyama method

**while** maximum amount of iterations not reached **do**

- Simulate a second Brownian Motion  $\bar{W}$  independently from  $W$  and set  $W^\circ = \rho W + \sqrt{1 - \rho^2} \bar{W}$ ;
- Simulate a solution to equation (4.7) with  $W^\circ$  as driving Brownian motion;
- Accept the proposal with probability  $\min \left\{ 1, \frac{\psi(T, X^\circ)}{\psi(T, X)} \right\}$  and set  $X = X^\circ$  and  $W = W^\circ$ .
- Repeat;

**end**

**Algorithm 1:** Algorithm for simulating a diffusion bridge using a given acceptance rate  $\rho$

In this algorithm, we start with a given solution and alter the driving Brownian motion via  $\rho$ . If we improve, we save the new solution and if we do not improve, we accept the new simulation with probability  $\frac{\psi(T, X^\circ)}{\psi(T, X)}$ .

**Remark 4.3.1.** Notice that algorithm 1 is a simplified version of algorithm 5.1 seen in van der Meulen and Schauer, 2017 [28] where the diffusion is conditioned on hitting multiple points with a given uncertainty. In lemma 5.2 of this paper, the convergence of this algorithm to the target distribution is proven. This shows that if we iterate long enough, the sampled paths have law  $\mathbb{P}_T^*$ .

**Remark 4.3.2** (Guided proposal stays on manifold). Notice that if  $X$  is a diffusion on a Riemannian manifold  $(\mathcal{M}, g)$ , then  $X^*$  stays on  $\mathcal{M}$  as well per definition, and hence, if we consider the set of functions that maps  $[0, T]$  to  $\mathcal{M}$  as  $A = \{f : [0, T] \rightarrow \mathbb{R}^N : f([0, T]) \subseteq \mathcal{M}\}$ , we must have  $\mathbb{P}_T^*(A) = \mathbb{P}(X_t^* \in \mathcal{M} \text{ for all } t \in [0, T]) = 1$ . Since  $\mathbb{P}_T^*$  and  $\mathbb{P}_T^\circ$  are equivalent, we also deduce that almost surely  $X^\circ$  stays on  $\mathcal{M}$  as well since  $\mathbb{P}(X_t^\circ \in \mathcal{M} \text{ for all } t \in [0, T]) = \mathbb{P}_T^\circ(A) = 1$ .

### 4.3.1. Choosing the auxiliary process

We have now seen that a guided proposal for a diffusion bridge is obtained by using the transition density of a linear process  $\tilde{X}$  governed by the SDE

$$d\tilde{X}_t = [\tilde{B}(t)\tilde{X}_t + \tilde{\beta}(t)] dt + \tilde{\sigma}(t) dW_t$$

where  $\tilde{\sigma}$  should be such that  $\tilde{\sigma}(T) = \sigma(T, v)$ . Equation (4.8) is satisfied as long as  $X^\circ$  behaves nicely as  $t \uparrow T$ . This means that there is some freedom left in choosing  $\tilde{B}$  and  $\tilde{\beta}$ . One possibility is to choose  $\tilde{B}$  and  $\tilde{\beta}$  equal to 0 and let  $\tilde{\sigma}$  be a constant function equal to  $\sigma(T, v)$ . This is a valid process such that equation (4.8) holds and  $\tilde{p}$  is easy to calculate as this would be the a scaled version transition density of Brownian motion. However, when we study appendix A of Bierkens et al., 2018 [3], we observe that existence of transition densities extends to control theory. Lemma A.2 tells us that the existence of a non-degenerate transition density is equivalent with complete controllability of the tuple  $(\tilde{B}, \tilde{\sigma})$ , which is in turn equivalent to  $\text{rank}(C) = d$ , where  $C$  is the controllability matrix given by

$$C := \begin{pmatrix} \tilde{\sigma} & B\tilde{\sigma} & \dots & \tilde{B}^{d-1}\tilde{\sigma} \end{pmatrix}.$$

In many applications  $\sigma(T, v)$  is not of full rank, so choosing  $\tilde{\sigma} = \sigma(T, v)$  constant and  $\tilde{B} = 0$  would imply that  $C$  is not of full rank. If this is the case, we opt to take  $\tilde{B}(t)_{i,j} = U_{ij}$ , where  $\{U_{ij}\}$  are drawn independently from an  $\mathcal{U}[0, 1]$  distribution. Alternative choices for the auxiliary process are suggested in section 5.1 of van der Meulen and Schauer, 2017b [28].

### 4.3.2. Numerical experiments of simulations of diffusion bridges

Let us start this section by demonstrating what a diffusion bridge should look like. We can do this by simulating a path that follows from equation (4.7) via the Euler-Maruyama method.

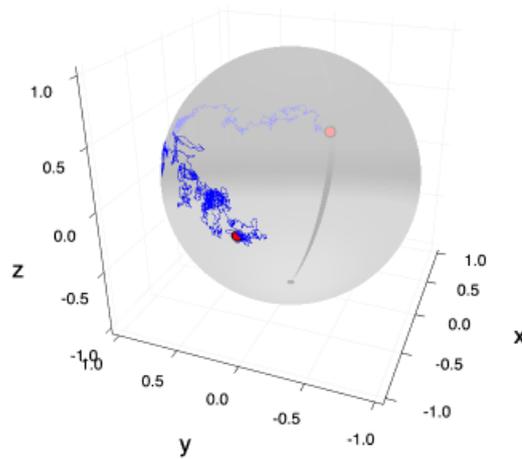


Figure 4.7: A simulated path of a diffusion bridge on a sphere that starts at  $(1, 0, 0)$  and is conditioned to hit  $(-1, 0, 0)$  at time  $T = \frac{1}{2}$ . The simulation was done by applying the Euler-Maruyama method to equation (4.7) with the drift and diffusion coefficient that correspond with the local coordinates as seen in example 3.1.2.

In order to improve on this first simulation, we can now apply algorithm 1. This is, however, when we observe singularities, as after some iterations, the likelihood explodes to the order of  $10^{72}$ . Figure 4.8 demonstrates the output paths in  $x, y, z$ -coordinates.

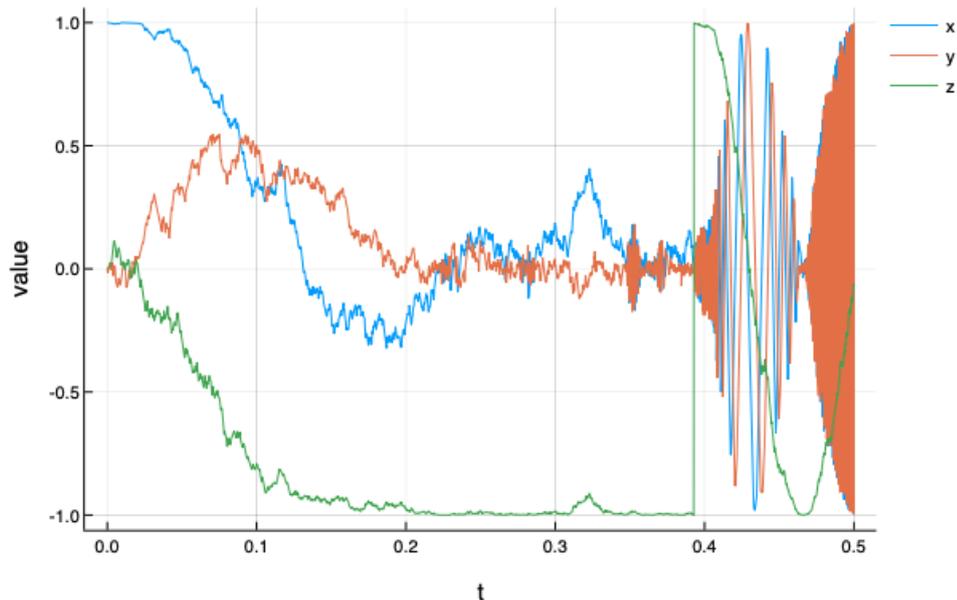


Figure 4.8: The  $x, y, z$ -coordinates after 50000 iterations of algorithm 1 applied on the sphere in local coordinates.

In order to explain the behavior that is observed in figure 4.8, we recall from example 3.1.2 that we simulate a diffusion bridge  $X^*$  that is obtained from conditioning a diffusion process process  $X$  on hitting  $(-1, 0, 0)$  at time  $T = \frac{1}{2}$ . The process  $X$  is governed by the SDE

$$dX_t = b(X_t) dt + \sigma(X_t) dW_t, \quad X_0 = (1, 0, 0)$$

where

$$b(x) = \begin{pmatrix} 0 \\ \frac{1}{2 \tan x^2} \end{pmatrix} \quad \text{and} \quad \sigma(x) = \begin{pmatrix} \frac{1}{|\sin x^2|} & 0 \\ 0 & 1 \end{pmatrix}$$

If we plug this into equation (4.10), we see that there are singularities occur when  $v \in \pi\mathbb{Z}$ . These points correspond to the  $x, y, z$ -coordinates  $(0, 0, 1)$  and  $(0, 0, -1)$ , points that are approached by our diffusion bridge after a sufficient amount of iterations of algorithm. If we simulate a bridge, say  $X^\circ$ , that approaches the north or south pole, we observe in equation (4.10), that the  $\psi(T, X^\circ)$  tends to  $\infty$ . We then observe in algorithm 1, that this bridge is accepted and from that point,  $\psi(T, X)$  tends to  $\infty$  and thus no new proposals will be accepted. Also note that the diffusion coefficient tends to  $\infty$  as  $V_t$  approaches  $k\pi$  for  $k \in \mathbb{Z}$ , which explains the behavior we see in figure 4.8. When simulating Brownian motion in local coordinates on the sphere using guided proposals, we thus encounter the problem that the most likely absolute continuity fails, which is caused by the fact that the drift and diffusion coefficient are unbounded.

# 5

## Likelihood-Based Inference on Riemannian manifolds

Let us revisit the aim of this thesis as stated in the introduction. We let  $X = (X_t)_t$  be a Brownian motion on a Riemannian manifold starting at  $\theta$  characterized in local coordinates or via orthogonal projections. In the model of this thesis, we have  $n$  observations  $\xi_1, \dots, \xi_n$  of  $X_T$ . As we have seen in section 2.3, the likelihood for  $\theta$  is given by

$$L(\theta | T, \xi) = \prod_{i=1}^n p(0, \theta; T, \xi_i)$$

Since  $p$  is generally not known in closed form, we have to approximate it. In this chapter, we elaborate on a stochastic approximation of  $p$  using diffusion bridges.

### 5.1. Approximating transition densities using diffusion bridges

We use the same notation as introduced in section 4.3, where we let  $X$  be a diffusion process,  $X^*$  the conditioned process  $X^* = (X | X_T = v)$  and  $X^\circ$  the guided proposal given in equation (4.7) with laws  $\mathbb{P}_T, \mathbb{P}_T^*$  and  $\mathbb{P}_T^\circ$  respectively. In section 4.3, we derived that under certain conditions

$$\frac{d\mathbb{P}_T^*}{d\mathbb{P}_T^\circ}(X^\circ) = \frac{\tilde{p}(0, \theta; T, v)}{p(0, \theta; T, v)} \psi(T, X^\circ) \quad (5.1)$$

Note that  $\frac{d\mathbb{P}_T^*}{d\mathbb{P}_T^\circ}$  is a function  $(\mathbb{R}^d)^{[0, T]} \rightarrow \mathbb{R}$ , so we can integrate both sides of equation (5.1) with respect to the measure  $\mathbb{P}_T^\circ$ . This yields

$$\int \frac{d\mathbb{P}_T^*}{d\mathbb{P}_T^\circ}(X^\circ) d\mathbb{P}_T^\circ(X^\circ) = \frac{\tilde{p}(0, \theta; T, v)}{p(0, \theta; T, v)} \int \psi(T, X^\circ) d\mathbb{P}_T^\circ(X^\circ)$$

where the left-hand side equals  $\int d\mathbb{P}_T^* = 1$ . We therefore observe that

$$p(0, \theta; T, v) = \tilde{p}(0, \theta; T, v) \mathbb{E} \psi(T, X^\circ) \quad (5.2)$$

With slight abuse of notation, we now let  $X_{(i)}^\circ$  denote the proposed diffusion bridge starting at  $\theta$  and ending at  $\xi_i$  and set  $X^\circ = (X_{(1)}^\circ, \dots, X_{(n)}^\circ)$  and  $\xi = (\xi_1, \dots, \xi_n)$ . Since we assume the data to be independent realizations, we obtain that the posterior density of  $\theta$  must satisfy

$$f(\theta | \xi, T) \propto \prod_{i=1}^n \tilde{p}(0, \theta; T, \xi_i) \mathbb{E} \psi(T, X_{(i)}^\circ)$$

and that the log-likelihood is given by

$$\ell(\theta | \xi, X^\circ) := \sum_{i=1}^n \log p(0, \theta; T, \xi_i) = \sum_{i=1}^n (\log \tilde{p}(0, \theta; T, \xi_i) + \log \mathbb{E} \psi(T, X_{(i)}^\circ)) \quad (5.3)$$

Note that, when estimating  $\mathbb{E}\psi(T, X^\circ)$  with  $\psi(T, x^\circ)$ , where  $x^\circ$  is a simulation of a sample path of  $X^\circ$  or with an average  $\frac{1}{m} \sum_{j=1}^m \psi(T, x_j^\circ)$  of simulations  $x_j^\circ$  of sample paths of  $X^\circ$ , the right hand side of equation (5.3) does not depend on any unknown variables and can thus be calculated using simulated diffusion bridges. In order to do calculations of the likelihood, we use that

$$\log \tilde{p}(0, \theta; T, v) = \tilde{R}(0, \theta)$$

In order to find  $\tilde{R}$ , we specify section 2.2 of van der Meulen and Schauer, 2017b [28] to our statistical model. Let us first define  $\Phi$  as the solution to

$$d\Phi(t) = \tilde{B}(t) \Phi(t) dt, \quad \Phi(0) = I$$

and set  $\Phi(t, s) = \Phi(t) \Phi(s)^{-1}$ . Furthermore, set

$$\tilde{L}(t) = \Phi(T, t) \quad \text{and} \quad \mu(t) = \int_t^T \tilde{L}(\tau) \tilde{\beta}(\tau) d\tau, \quad t \in [0, T].$$

**Theorem 5.1.1.** *Assume that for  $t \in [0, T]$ , the matrix*

$$\tilde{M}(t) = \left( \int_t^T \tilde{L}(\tau) \tilde{\alpha}(\tau) \tilde{L}(\tau)^T d\tau \right)^{-1}$$

*exists. Then*

$$\tilde{r}(t, x) = \tilde{L}(t)^T \tilde{M}(t) (v - \mu(t) - \tilde{L}(t)x) .$$

*Proof.* This is a simplified version of theorem 2.3 of van der Meulen and Schauer, 2017b [28] □

All that is left to find  $\tilde{R}$  is to note that  $\tilde{R}$  satisfies  $\tilde{r} = \nabla_x \tilde{R}$  at  $(t, x)$  when

$$\tilde{R}(t, x) = \log \varphi(v; \mu(t) + \tilde{L}(t)x, \tilde{M}(t)^{-1}) \tag{5.4}$$

where  $\varphi(\cdot; \mu, \Sigma)$  denotes the density of the multivariate normal distribution with mean  $\mu$  and covariance matrix  $\Sigma$ .

In equation (5.3), we can now substitute equation (5.4) for  $\log \tilde{p}$  and approximate  $\log \mathbb{E}\psi(T, X^\circ)$  via simulations of diffusion bridges to find a stochastic approximation for the log-likelihood of  $\theta$ .

## 5.2. Drawing samples from the likelihood of $\theta$

In order to draw samples from the likelihood of  $\theta$ , we use Markov chain Monte Carlo (MCMC) methods, see appendix B for more information on these algorithms.

In order to specify the algorithms, we introduce the shorthand notation  $GP(\theta, W, v)$  for a guided proposal for a diffusion bridge on  $\mathcal{M}$  starting at  $\theta$ , conditioned on hitting  $v$  and with driving Brownian motion. Let  $q(\theta, \cdot)$  be the distribution function for drawing a proposed update  $t$ . We use Metropolis-Hastings algorithm 2 for drawing samples from the likelihood while simultaneously updating the diffusion bridges between the data point and  $\theta$ .

This algorithm consists of two steps. In the first step, we update the guided proposals in order to increase the value of  $\frac{d\mathbb{P}^*}{d\mathbb{P}^\circ}(X^\circ)$ . This ensures that we draw diffusion bridges whose laws get closer to  $\mathbb{P}^*$ , the true law of the conditioned process. Step 1 therefore ensures that we get better approximations to the log-likelihood of  $\theta$ . In the second step we propose a new  $\theta$  and either accept or reject. This step results in a draw from the likelihood after a sufficient amount of iterations.

**Input:** Start with random point  $\theta \in \mathcal{M}$  and tuning parameters  $\varepsilon$  and  $\rho > 0$ . Furthermore, simulate paths standard Brownian motions  $W_{(1)}, \dots, W_{(n)}$  in  $\mathbb{R}^d$  and diffusion bridges  $X_{(1)}, \dots, X_{(n)}$  where  $X_{(i)} = GP(\theta, W_{(i)}, \xi_i)$ .

**while** maximum number of iterations not reached **do**

**Step 1:** Update the diffusion bridges and driving Brownian motions.

**for all data points**  $\xi_i$  **do**

Simulate a Brownian motion  $\bar{W}$  in  $\mathbb{R}^d$  independent of  $W_{(i)}$  and set

$$W_{(i)}^\circ = \rho W_{(i)} + \sqrt{1 - \rho^2} \bar{W};$$

Simulate a diffusion bridge  $X_{(i)}^\circ = GP(\theta, W_{(i)}^\circ, \xi_i)$ ;

With probability  $\min\left\{1, \frac{\psi(T, X_{(i)}^\circ)}{\psi(T, X_{(i)})}\right\}$ , accept the proposal and set  $X_{(i)} = X_{(i)}^\circ$  and

$$W_{(i)} = W_{(i)}^\circ;$$

**end**

**Step 2:** Update the parameter  $\theta$ .

Propose a new point  $\theta^\circ$  from the distribution  $q(\theta, \cdot)$ ;

**for all data points**  $\xi_i$  **do**

Simulate a bridge  $X_{(i)}^{\circ\circ} = GP(\theta^\circ, W_{(i)}, \xi_i)$ ;

**end**

With probability  $\min\left\{1, \frac{L(\theta^\circ | \xi, X^{\circ\circ}) q(\theta^\circ, \theta)}{L(\theta | \xi, X) q(\theta, \theta^\circ)}\right\}$ , where  $L = \exp \ell$  and  $\ell$  is the stochastic approximation to equation (5.3), accept the proposed update for  $\theta$  and set  $\theta = \theta^\circ$  and  $X = X^{\circ\circ}$ ;

**end**

**Algorithm 2:** Metropolis-Hastings algorithm for sampling from the likelihood of  $\theta$ .

### 5.2.1. Uniform proposals

The first algorithm we propose is the MH algorithm 2 with a uniform proposal distribution for  $\theta$ . Here the notion of a uniform proposal is to be interpreted as follows. We consider a uniform update to the local coordinates, since they are homeomorphic to  $\mathbb{R}^d$  and we can thus add a uniform draw to each of the local coordinates. An advantage of these proposals is that they are symmetric, that is  $q(x, y) = q(y, x)$  and therefore this term vanishes when calculating the acceptance probability for the updates on  $\theta$ . A disadvantage is that, since the updates are uniform and therefore do not use any knowledge of the data, it takes more iterations before we are sampling from the likelihood of  $\theta$ .

### 5.2.2. Langevin adjusted proposals

In the previous section we choose uniform proposals for the updates on  $\theta$ . In this section we aim to improve on these proposals by combining a gradient-based method with the Metropolis-Hastings algorithm. For some background theory on Langevin adjusted proposals, we refer to appendix B.2.1

When applying Langevin adjusted proposals, we draw our proposed updates  $\theta^\circ$  when we are at  $\theta$  from an  $\mathcal{N}(\theta + \frac{h}{2} \nabla \ell(\theta), hI)$ -distribution. This method has the advantage that each update moves in the direction of  $\nabla \ell$  and therefore moves towards the maximum of  $\ell$ . This means that fewer iterations are needed in order to draw from the likelihood of  $\theta$ .

## 5.3. Maximum likelihood estimates

It of interest to see if we can also find maximum likelihood estimates. This maximum can be found using gradient-based methods. The idea behind this method is to start at a point  $\theta_0 \in \mathcal{M}$  and move to a new point  $\theta_1$  in the direction of the gradient of  $\ell$  at  $\theta_0$ .

A standard gradient ascent method proposes an update  $\theta_1 = \theta_0 + h \nabla \ell(\theta_0)$  for a specified tuning parameter  $h > 0$ . Under certain regularity conditions, these updates converge to the maximum of  $\ell$ . We however, have the problem that  $\ell$  is intractable and we move to the stochastic approximation using diffusion bridges, a method we refer to as stochastic gradient ascent.

A problem encountered when applying stochastic gradient ascent on a manifold is that the updates generally do not lie on the manifold. Since the gradient  $\nabla\ell(\theta_0)$  defines the direction of the updates, we propose to project the gradient to the tangent space  $T_{\theta_0}\mathcal{M}$  via the orthogonal projection matrix  $P(\theta_0)$  we observed in section 3.2 and apply the exponential map if known in closed form the the tangent vector we obtain after projecting  $\nabla\ell(\theta_0)$ . This results in algorithm 3.

```

Input: Start with random point  $\theta_0$  on  $\mathcal{M}$ .
Set  $\theta_1 = \text{Exp}_{\theta_0}(hP(\theta_0)\nabla\ell(\theta_0))$ ;
if  $\|\theta_0 - \theta_1\| < \varepsilon$  then
  | Stop
else
  | Set  $\theta_0 = \theta_1$  and repeat
end

```

**Algorithm 3:** Algorithm for finding a maximum likelihood estimate using stochastic gradient descent in combination with the exponential map.

Notice that when the updates gets closed to the maximum, the norm of  $\nabla\ell$  decreases. Therefore the geodesic distance (see 2.1.32) between the updates goes to 0, which demonstrates that this algorithm indeed converges.

For the implementation of both the gradient-based method and the Langevin adjusted proposals, we use the ForwardDiff.jl package [18] in Julia. Here, the gradient can be approximated quickly by using automatic differentiation. For the simulations of diffusions, we can make use of the Bridge.jl package by Schauer et al., 2018 [23].

# 6

## Numerical experiments on the sphere

In order to test the theory derived in chapter 5 on the calculation of log-likelihoods, we generated a dataset of size 100 on the sphere by simulating independent Brownian motions on the sphere and saving their values at a fixed time  $T$ .

Since we have seen that a characterization of Brownian motion using charts can be problematic when simulating bridges on the sphere, we choose the characterization via the Stratonovich SDE  $dX_t = P(X_t) \circ dW_t$ . For the simulations we use the Euler-Maruyama method, as there is no straightforward way for finding 2-jets that characterize the SDEs for Brownian motion and bridges on the sphere.

### 6.1. Deriving the Itô SDE and calculations on a grid of points

We first calculate the projection matrix on the sphere using corollary 3.2.5. For any point  $x$  on the sphere, the normal vector to the point is given by  $n(x) = x$  and we thus have that

$$P(x) = I - n(x)n(x)^T = \begin{pmatrix} 1 - (x^1)^2 & -x^1x^2 & -x^1x^3 \\ -x^1x^2 & 1 - (x^2)^2 & -x^2x^3 \\ -x^1x^3 & -x^2x^3 & 1 - (x^3)^2 \end{pmatrix}.$$

Now recall that, when applying the Euler-Maruyama method to an SDE, it is assumed that the SDE is in Itô form and thus we translate it into this form first. A direct computation, using that  $(x^1)^2 + (x^2)^2 + (x^3)^2 = 1$  yields

$$\frac{1}{2} \sum_{j,k} P(x)_{jk} \partial_j \partial_k P(x)_{ik} = -x^i, \quad i = 1, 2, 3$$

and we may thus conclude that the Stratonovich SDE  $dX_t = P(X_t) \circ dW_t$  coincides with the Itô SDE given by

$$dX_t = -X_t dt + P(X_t) dW_t. \quad (6.1)$$

We apply the Euler-Maruyama method to simulate diffusions starting at a given point and save them at a fixed time  $T$  to generate a dataset on the sphere that coincides with our model. Figure 6.1 shows the simulated dataset when taking  $\theta = (0, 0, 1)$  as starting point and  $T = \frac{1}{2}$ .

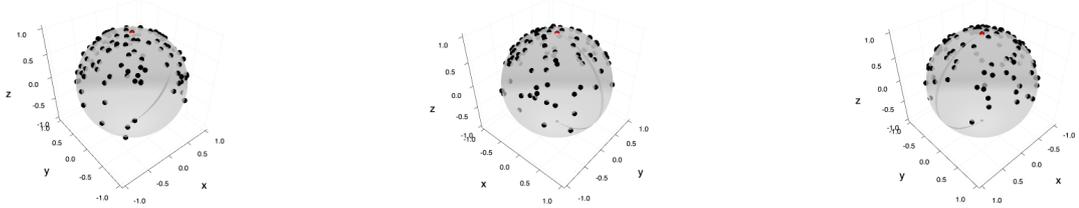


Figure 6.1: A view from three separate angles on a dataset on the sphere simulated by saving the values of simulations of realizations of the solution to equation (6.1) using the Euler-Maruyama method. The red dot indicates the point  $\theta = (0, 0, 1)$  that was used as the starting point for the diffusions.

We can now apply the theory derived in chapter 5, to approximate the log-likelihood on a grid of points in order to get an impression of how this approximation performs. We thus create a grid of points (of size 5200) and at each point, we use the Euler-Maruyama method to simulate guided proposals for diffusion bridges between that grid point and each data point. These bridges can then be used in equation (5.3) to approximate the log-likelihood at that grid point. The results are as given in figure 6.2.

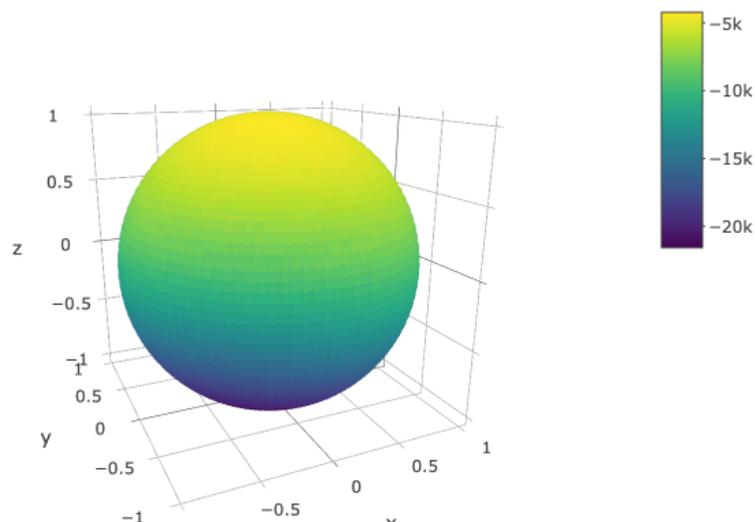


Figure 6.2: The log-likelihood as stated in equation (5.3) calculated at a grid of points on the sphere by simulating diffusion bridges between these points and the data points simulated previously.

Figure 6.2 demonstrates what we expect. Recall that data was simulated from Brownian motion on the sphere starting at  $\theta = (0, 0, 1)$  and we see that our log-likelihood increases as we move close to the north pole, whereas it decreases when we move away from it. The point on the grid where the log-likelihood attains its maximum is given by the point  $(-0.278, -0.027, 0.960)$ .

## 6.2. Finding the maximum likelihood estimator

Calculating the log-likelihood in each point on a small grid is computationally feasible in this example, but in practice, it would be useful to test the gradient-based method derived in section 5.3 to find a maximum likelihood estimate for  $\theta$ .

### 6.2.1. Stochastic gradient descent using the exponential map

On the sphere, the exponential map is known in closed form (see corollary 2.1.35) and we can thus apply algorithm 3 to this manifold.

We apply this algorithm many times with many different starting points on the sphere and When typically observe a maximum at  $(-0.278, 0.0289, 0.9603)$ , which coincides with the maximum found on our grid of figure 6.2 earlier. The amount of iterations needed to find this maximum naturally depends on the random starting point and the convergence rate  $h$ . After some trial and error, we observed convergence after an amount of iterations that was typically between 30 and 50 when  $h$  is of the order  $10^{-4}$  and we stop the algorithm when the updates are within a distance of  $10^{-6}$  from each other. A somewhat more graphical representation of this result can be seen in figure 6.3, where we plotted the  $x, y, z$ -components of the updates at each iteration of one of the runs algorithm 3.

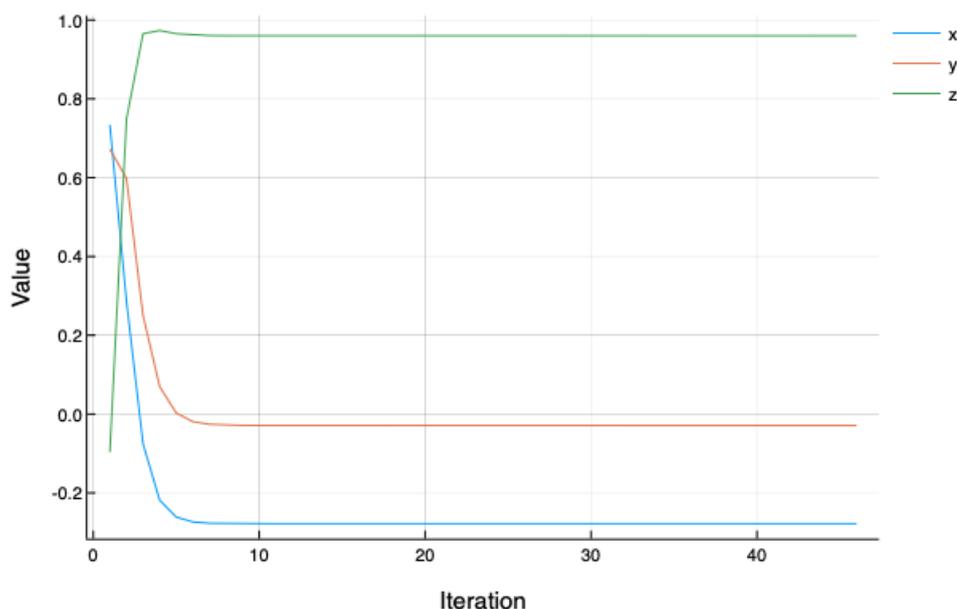


Figure 6.3: The  $x, y, z$ -components at each iteration of the gradient descent algorithm that uses the exponential map for the updates

Figure 6.3 demonstrates that there is convergence present to the maximum likelihood estimate after a small amount of iterations. After some iterations, the updates get closer to each other and at after a sufficient amount of iterations, the distance between the updates reaches the threshold set at  $10^{-6}$  and then the algorithm stops. Figure 6.4 demonstrates this behavior in the trajectory of the updates of the algorithm plotted in  $\mathbb{R}^3$ .

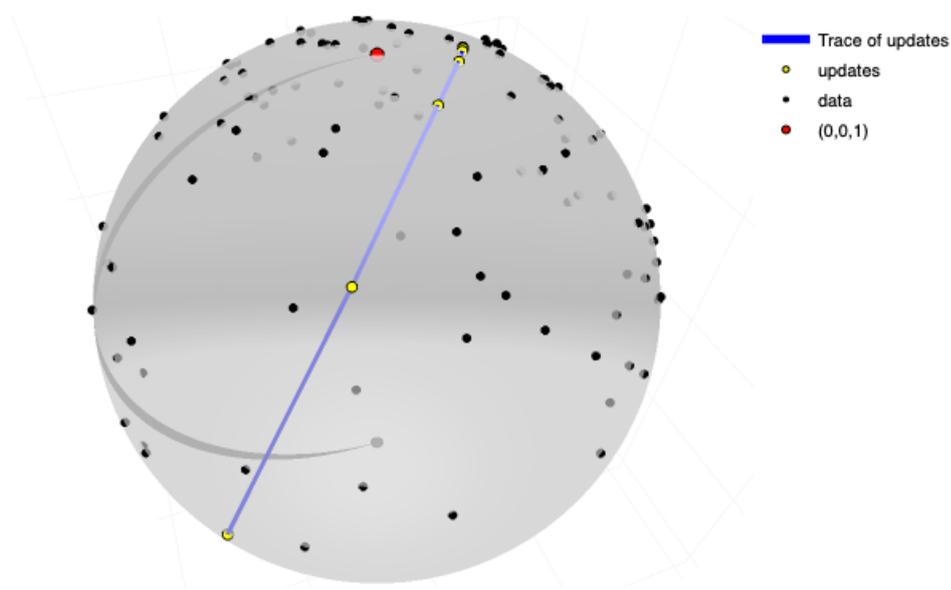


Figure 6.4: The trajectory of the updates of the gradient descent algorithm 3 plotted in  $\mathbb{R}^3$  (blue) with the data points added to the figure (black) as well as the north pole  $(0, 0, 1)$  from which the dataset was simulated.

### 6.2.2. Stochastic gradient descent while normalizing the updates

Here, we use the exponential map in combination with the orthogonal projection of the gradient in order to guarantee that the updates remain on the sphere. It is interesting to see what happens when we use the specific property of the sphere that we can project any point in  $\mathbb{R}^d$  onto it by normalizing to find out if we can improve upon algorithm 3. We are now able to move our updates directly in the direction of the gradient of  $\ell$  and normalize the result. This results in algorithm 4.

**Input:** Start with random point  $\theta_0$  on the sphere.

Set  $\theta_1 = \frac{\theta_0 + h \nabla \ell(\theta_0)}{\|\theta_0 + h \nabla \ell(\theta_0)\|}$ ;

if  $\|\theta_0 - \theta_1\| < \varepsilon$  then

  | Stop

else

  | Set  $\theta_0 = \theta_1$  and repeat

end

**Algorithm 4:** Algorithm for finding a maximum likelihood estimate using stochastic gradient descent while normalizing the updates.

The results of algorithm 4 are similar to the results of algorithm 3; after about 30 to 60 iterations, the algorithm converges. In this instance, the convergence is to the point  $(-0.278, -0.0289, 0.960)$ , which is again close to the maximum on the grid and the maximum found using the exponential map. In figure 6.5, we show a similar plot to the exponential map of the separate  $x, y, z$ -components when this method is used.

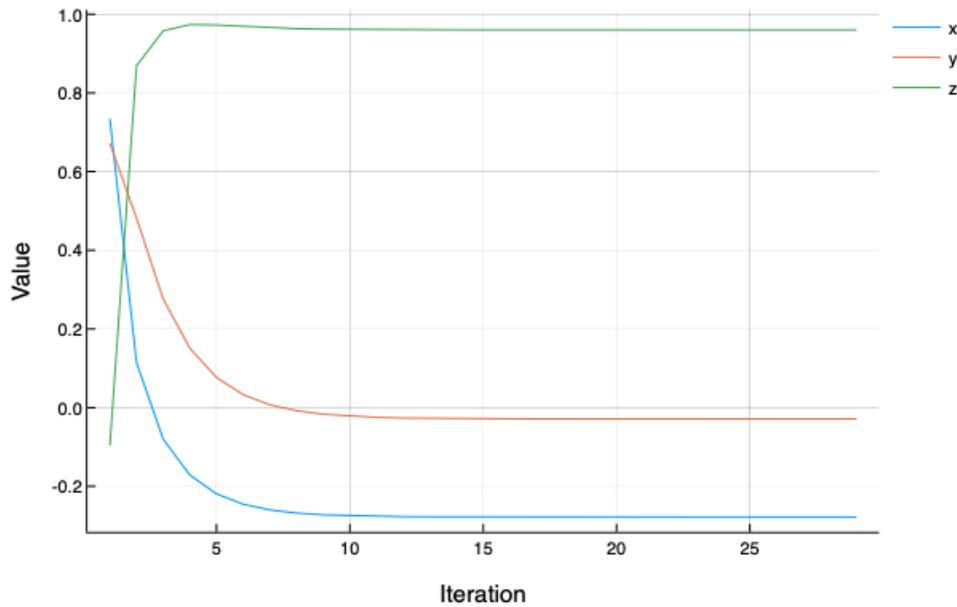


Figure 6.5: The  $x, y, z$ -components at each iteration of the gradient descent algorithm that normalizes the updates

We see the same behavior in figure 6.5 as we saw in figure 6.3, after about 10 iterations, we get close to our maximum likelihood estimate and then it takes more iterations to reach the threshold of  $10^{-6}$  before the algorithm is stopped. Figure 6.6 also demonstrates this behavior in the trajectory of the updates of the algorithm on the sphere.

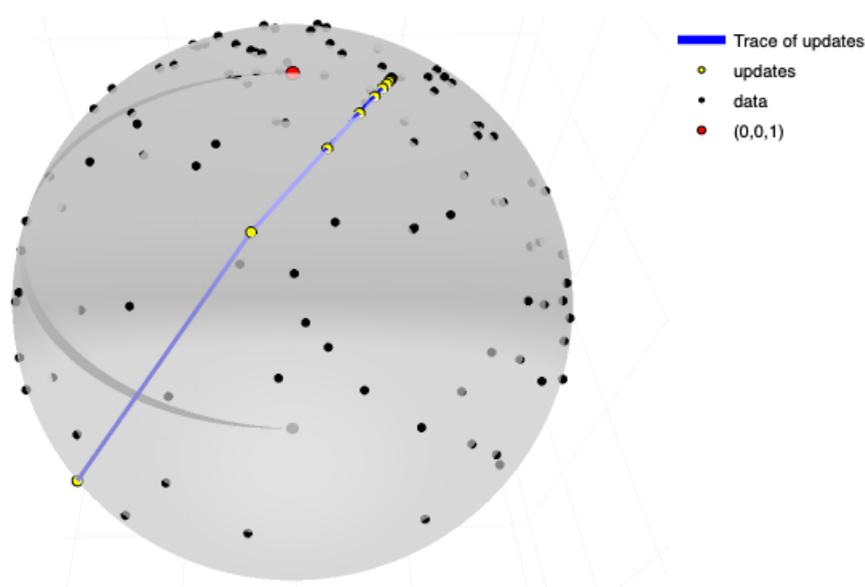


Figure 6.6: The trajectory of the updates of the gradient descent algorithm 3 plotted in  $\mathbb{R}^3$  (blue) with the data points added to the figure (black) as well as the north pole  $(0, 0, 1)$  from which the dataset was simulated.

### 6.3. Markov Chain Monte Carlo methods

We now apply the Markov chain Monte Carlo methods derived in section 5.2 to draw samples from the likelihood of  $\theta$ .

### 6.3.1. Uniform proposals

We first translate algorithm 2 for uniform proposals. Uniform proposals can now be given as follows. Given a point  $\theta = (\cos \vartheta \sin \varphi, \sin \vartheta \sin \varphi, \cos \varphi)$ , we propose uniform updates given by  $\theta^\circ = (\cos(\vartheta + U_1) \sin(\varphi + U_2), \sin(\vartheta + U_1) \sin(\varphi + U_2), \cos(\varphi + U_2))$  where  $U_1, U_2 \sim \mathcal{U}(-\varepsilon, \varepsilon)$ . We applied this method to the same dataset with  $\varepsilon = \frac{\pi}{100}$  and  $\rho = \frac{1}{2}$  to find the stationary distributions for the  $x, y, z$ -coordinates given in figure 6.7.

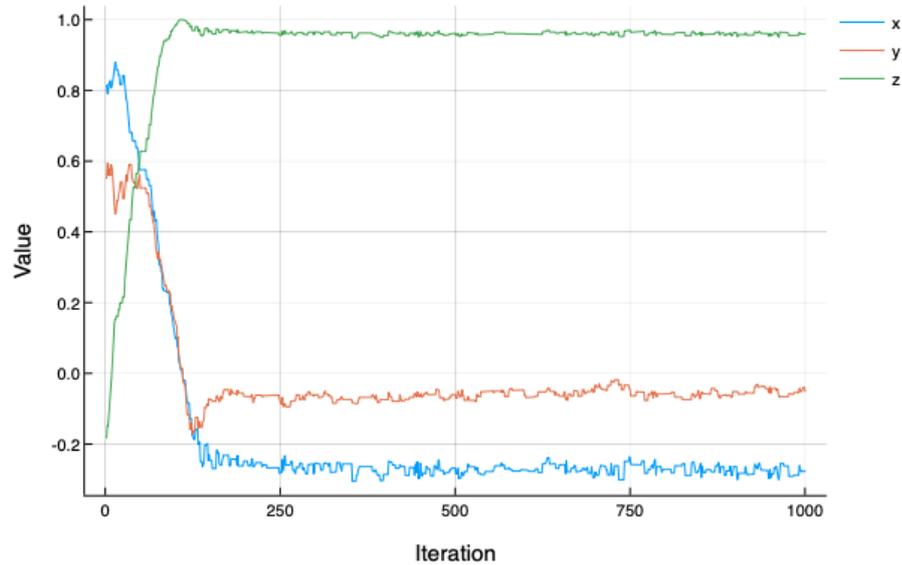


Figure 6.7: The resulting Markov chain after applying the Markov Chain Monte Carlo algorithm 2 with uniform proposals to the dataset of size 100 simulated with starting point  $(0, 0, 1)$ .

In figure 6.7, we see a convergence to the stationary distribution. Recall that the optimum on a grid of size 5200 was found at  $(-0.278, -0.027, 0.960)$ . The Markov chain generated is stable around the same point. A multidimensional visualization of this result can be found in figure 6.8.

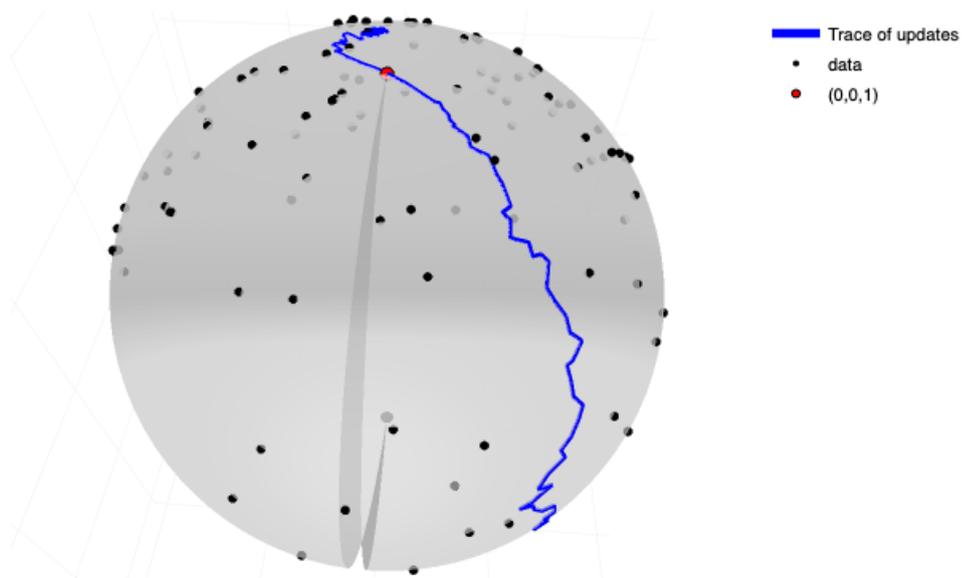


Figure 6.8: The Markov chain resulting from applying algorithm 2 with uniform proposals plotted in  $\mathbb{R}^3$  (blue) with the data points added to the figure (black) as well as the north pole  $(0, 0, 1)$  from which the dataset was simulated.

Figure 6.8 more clearly demonstrates what we already saw in figure 6.7, where we saw that the Markov chain converged to its stationary distribution. We can see this here when noticing that the Markov Chain stays around the same point after a few iterations.

### 6.3.2. Langevin adjusted proposals

In this section, we apply the Langevin adjusted proposals as seen in section 5.2.2. We do however make a slight alteration. In order to guarantee that the updates stay on the sphere, we propose updates given by

$$\theta^o = \frac{\theta + \frac{h}{2} \nabla \ell(\theta) + \sqrt{h} Z}{\left\| \theta + \frac{h}{2} \nabla \ell(\theta) + \sqrt{h} Z \right\|}$$

When we apply algorithm 2 with this proposal distribution to the same simulated dataset with  $\rho = \frac{1}{2}$  and  $h = 10^{-4}$ , we typically observe convergence within 20 iterations. One of the traces of updates can be seen in figure 6.9.

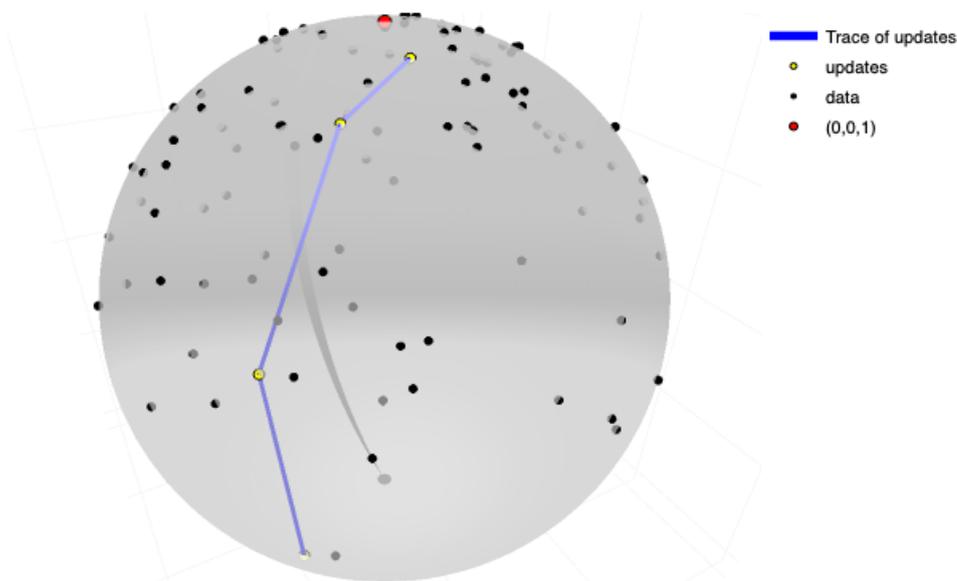


Figure 6.9: A trace of the updates resulting from applying algorithm 2 with Langevin adjusted proposals to the dataset simulated from  $(0, 0, 1)$ . The data points are added to the plot.

When looking at figure 6.9, we notice that the Langevin updates behave very similar to the gradient-based methods for optimizing. The reason for this is that in this case, the norm of the gradient of the log-likelihood is of the order  $10^5$ . This causes the normal random variable in equation (B.1) to become negligible compared to the gradient term and thus eliminates the randomness in the updates.

## 6.4. Notes

In this chapter, we studied the main statistical problem of this thesis in case the manifold  $\mathcal{M}$  is a sphere. Figure 6.2 demonstrated that the method derived for approximating the log-likelihood stated in equation (5.3) behaves as we expect, as we saw high values around the north pole and low values around the south pole. The various methods of finding maxima all have their advantages and disadvantages. The gradient based methods converged quickly to one single point that coincides with the maximum on the grid. A disadvantage of the gradient-based methods is that they strongly depend on the structure of the sphere, where the exponential map is known in closed form and projecting is equivalent to normalizing. Such methods are therefore difficult to apply to other manifolds. MCMC methods on the other hand have the advantage that they can be combined with the calculation of the likelihood, as we can simultaneously update the diffusion bridges between our points and move on the sphere to a maximum.

We conclude this chapter by noting that none of the algorithms did give a maximum of  $(0, 0, 1)$ . The most likely reason for this is the sample size of 100 in combination with the discretization error made when simulating the diffusion bridges.

# 7

## Conclusion

We conclude this thesis by going through the broad lines of work that we have done. The aim was to do likelihood-based inference on the center of a dataset with values on a Riemannian manifold and in order to get there, we started off by studying Riemannian geometry before moving to diffusions in general as well as their transition density. The chapters after this focused on several ways to characterize Brownian motion on a manifold and several ways of simulating diffusions before demonstrating some results on the sphere and other nonlinear spaces.

### 7.1. Characterizing Brownian motion on manifolds

We have researched two main characterizations of Brownian motion on manifolds in this thesis; Brownian motion on local coordinates and Brownian motion via projections to the tangent space. Numerically, the characterization via local coordinates should outperform the method involving projections, as it requires less driving Brownian motions and does its calculations in a lower dimension. However, we have seen that problems arise when the simulations are approaching the boundaries of the charts. Armstrong and Brigo, 2018 [1] did propose a method that involves switching charts when one nears the boundary, so this is interesting to investigate further.

### 7.2. Simulation of diffusions

We also investigated two types of simulating diffusions, the Euler-Maruyama method and the method involving 2-jets. When a diffusion is given in the form a stochastic differential equation, the latter did require us to know the 2-jet from which fitted that SDE. Since the SDEs for Brownian motion on manifold were not trivial on itself and had to be manipulated to generate guided proposals for diffusion bridges, it was not practical to implement the 2-jet method in numerical experiments.

### 7.3. Numerical experiments

After deriving a method for estimating the log-likelihood of the starting point of Brownian motion, we did numerical experiments to test the result. The Bridge.jl package [23] in Julia provided us with an excellent tool to test our research and we have seen on the sphere that our approach to do likelihood-based inference behaves as we expect. However, the properties of the sphere were frequently used, which complicated experiments on other manifolds. Extending the gradient-based methods to general manifolds is therefore interesting future research.

### 7.4. Suggestions for future research

#### 7.4.1. Extend the methods to other spaces

In chapter 6, we have seen that gradient-based methods performed well for finding maximum likelihood estimates, but did rely heavily on the structure of the sphere. It is interesting to research if there is a more general method for gradient-based methods on manifolds. The MCMC method with uniform

proposals seems to be more easily translated to other spaces as it only relied on the local coordinates of the sphere.

#### 7.4.2. Weaken conditions for existence of transition densities

We have discussed global Lipschitz conditions on SDEs to guarantee existence of a transition density of the solution. When translating this to manifolds, we see that there are various conditions on the manifold for these conditions to apply. It is interesting to see if these conditions can be weakened. One could research if Hörmanders theorem can be applied to this problem.

#### 7.4.3. Other approximations of the the transition density $p$

In this thesis, we focused on a stochastic approximation to the transition density  $p$  of a diffusion process using diffusion bridges, but there are other ways to obtain an approximation. Interesting research would be to find out if it is possible to apply finite element or finite difference methods to the Fokker-Planck equation (2.39) in order to solve the PDE numerically. It is interesting research to make a comparison between classical numerical methods to approximate  $p$  and the stochastic approximation made in this thesis.

#### 7.4.4. Relevance of the parameter $T$

This thesis assumed a model where a set of data points is interpreted as hitting points of realizations of Brownian motion at a given time  $T$ . In the numerical experiments,  $T$  was fixed and the same  $T$  was used for simulating the data as well as making estimations for the center of the data. In practice, the value of  $T$  is unknown. Since a Brownian motion at time  $t$  has variance  $t$ , we suggest that  $T$  is also a parameter that is involved when estimating the variance in the dataset. If we look at classical statistics, where the mean and variance of a dataset can be estimated separately, we argue that this holds in our case as well and that fixing  $T$  was a valid assumption. Investigating the relevance and possible estimates of this parameter would yet be very interesting for future experiments.

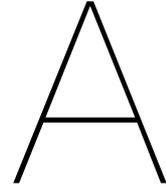
#### 7.4.5. Choices for the auxiliary process

In section 4.3.1, we discussed the auxiliary process that is used to generate guided proposals for diffusion bridges. As stated in this section, there is a lot of freedom in this choice and we have chosen a random matrix  $\tilde{B}$  to ensure controllability. However the question arises whether or not there is an optimal choice for the auxiliary process that ensures a high acceptance probability in the MH algorithms.

# Bibliography

- [1] J. Armstrong and D. Brigo. Intrinsic stochastic differential equations as jets. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2210):20170559, 2018. doi: 10.1098/rspa.2017.0559. URL <https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2017.0559>.
- [2] Krishna B Athreya, Hani Doss, Jayaram Sethuraman, et al. On the convergence of the markov chain simulation method. *The Annals of Statistics*, 24(1):69–100, 1996.
- [3] Joris Bierkens, Frank van der Meulen, and Moritz Schauer. Simulation of elliptic and hypo-elliptic conditional diffusions. *arXiv preprint arXiv:1810.01761*, 2018.
- [4] Alberto Bressan. *Lecture Notes on Functional Analysis: With Applications to Linear Partial Differential Equations*, volume 143. American Mathematical Soc., 2012. ISBN 978-0-8218-8771-4. URL [www.ams.org/bookpages/gsm143](http://www.ams.org/bookpages/gsm143).
- [5] Samuel N Cohen and Robert James Elliott. *Stochastic calculus and applications*, volume 2. Springer, 2015.
- [6] Bernard Delyon and Ying Hu. Simulation of conditioned diffusion and application to parameter estimation. *Stochastic Processes and their Applications*, 116(11):1660–1675, 2006.
- [7] Andreas Eberle. *Introduction to Stochastic Analysis*. 2016.
- [8] Desmond J Higham, Xuerong Mao, and Andrew M Stuart. Strong convergence of euler-type methods for nonlinear stochastic differential equations. *SIAM Journal on Numerical Analysis*, 40(3):1041–1063, 2002.
- [9] Elton P Hsu. *Stochastic analysis on manifolds*, volume 38. American Mathematical Soc., 2002. ISBN 978-0-8218-0802-3. doi: <http://dx.doi.org/10.1090/gsm/038>.
- [10] Olav Kallenberg. *Foundations of modern probability*. Springer Science & Business Media, 2006.
- [11] Rajeeva L Karandikar and BV Rao. On quadratic variation of martingales. *Proceedings-Mathematical Sciences*, 124(3):457–469, 2014.
- [12] Jean-François Le Gall. *Brownian motion, martingales, and stochastic calculus*, volume 274. Springer, 2016.
- [13] John M Lee. Smooth manifolds. In *Introduction to Smooth Manifolds*. Springer, 2013.
- [14] Stephen Lovett. *Differential geometry of manifolds*. AK Peters/CRC Press, 2010. ISBN 978-1-4398-6546-0.
- [15] TJ Lyons and WA Zheng. On conditional diffusion processes. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics*, 115(3-4):243–255, 1990.
- [16] Théo Michelot, Marie-Pierre Etienne, and Pierre Gloaguen. The langevin diffusion as a continuous-time model of animal movement and habitat selection. *arXiv preprint arXiv:1810.10213*, 2018.
- [17] Bernt Karsten Øksendal. *Stochastic Differential Equations: An Introduction With Applications*. Springer-Verlag, 2003.
- [18] J. Revels, M. Lubin, and T. Papamarkou. Forward-mode automatic differentiation in julia. *arXiv:1607.07892 [cs.MS]*, 2016. URL <https://arxiv.org/abs/1607.07892>.

- [19] Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.
- [20] L. C. G. Rogers and David Williams. *STOCHASTIC DIFFERENTIAL EQUATIONS AND DIFFUSIONS*, volume 2 of *Cambridge Mathematical Library*, page 110–303. Cambridge University Press, 2 edition, 2000. doi: 10.1017/CBO9780511805141.003.
- [21] Jeffrey S Rosenthal. *A first look at rigorous probability theory*. World Scientific Publishing Company, 2006.
- [22] Moritz Schauer, Frank van der Meulen, and Harry van Zanten. Guided proposals for simulating multi-dimensional diffusion bridges. *Bernoulli*, 23(4A):2917–2950, 11 2017a. doi: 10.3150/16-BEJ833. URL <https://doi.org/10.3150/16-BEJ833>.
- [23] Moritz Schauer et al. Bridge 0.9.0, 2018. URL <https://github.com/mschauer/Bridge.jl>.
- [24] Flora Spieksma and Harry van Zanten. *An Introduction to Stochastic Processes in Continuous Time: the non-Jip-and-Janneke-language approach*. 2016. URL <http://pub.math.leidenuniv.nl/~spieksmaf/colleges/sp-master/sp-hvz1.pdf>.
- [25] Michael D Spivak. *A comprehensive introduction to differential geometry*. Publish or perish, 1970. ISBN 0-914098-70-5.
- [26] Michael D Spivak. *Differential Geometry, Volume II, third edition*. Publish or perish, 1999. ISBN 0-914098-71-3.
- [27] M Van Den Berg and JT Lewis. Brownian motion on a hypersurface. *Bulletin of the London Mathematical Society*, 17(2):144–150, 1985.
- [28] Frank van der Meulen and Moritz Schauer. Continuous-discrete smoothing of diffusions. *arXiv e-prints*, art. arXiv:1712.03807, Dec 2017b. URL <https://ui.adsabs.harvard.edu/abs/2017arXiv171203807v>. Provided by the SAO/NASA Astrophysics Data System.



# Additional Theory on Martingales and stopping times

This appendix is included to provide some background material on martingales and formulate some useful results in this field of probability theory. Throughout this chapter we will assume that we have a filtered probability space  $(\Omega, \mathcal{F}, (\mathcal{F}_t)_t, \mathbb{P})$  that satisfies the usual conditions as stated in section 2.2

## A.1. Conditional expectation

Before being able to introduce martingales, we first need to introduce conditional expectations in a proper manner.

**Definition A.1.1** (Conditional expectation). *Let  $X : \Omega \rightarrow \mathbb{R}^d$  be a random variable and  $\mathcal{G} \subseteq \mathcal{F}$  a sub- $\sigma$ -field of  $\mathcal{F}$ . The conditional expectation of  $X$  given  $\mathcal{G}$  is any  $\mathcal{G}$ -measurable random variable  $\mathbb{E}(X | \mathcal{G}) : \Omega \rightarrow \mathbb{R}^d$  such that*

$$\int_G \mathbb{E}(X | \mathcal{G}) \, d\mathbb{P} = \int_G X \, d\mathbb{P}$$

for all  $G \in \mathcal{G}$

Note that this definition introduces a random variable that need not necessarily exist. However a proof of existence and uniqueness up to a set of measure 0 can for example be found in Rosenthal, 2006 [21].

**Theorem A.1.2** (Properties of conditional expectations). *Let  $X, Y$  be random variables and let  $\mathcal{G}, \mathcal{G}_1, \mathcal{G}_2$  be sub- $\sigma$ -fields of  $\mathcal{F}$ . Then the following properties hold.*

- (i) *Conditional expectation defines a linear operator in the sense that  $\mathbb{E}(\alpha X + \beta Y | \mathcal{G}) = \alpha \mathbb{E}(X | \mathcal{G}) + \beta \mathbb{E}(Y | \mathcal{G})$  for all  $\alpha, \beta \in \mathbb{R}^d$ .*
- (ii) *If  $X$  is independent of  $\mathcal{G}$ , i.e. the events  $\{X \in B\}$  and  $G$  are independent events for all Borel sets  $B$  and  $G \in \mathcal{G}$ , then  $\mathbb{E}(X | \mathcal{G}) = \mathbb{E}X$ .*
- (iii) *If  $X$  is  $\mathcal{G}$ -measurable, then  $\mathbb{E}(X | \mathcal{G}) = X$ , furthermore if  $Y$  is another random variable, then  $\mathbb{E}(XY | \mathcal{G}) = X \mathbb{E}(Y | \mathcal{G})$ .*
- (iv) *Suppose that  $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \mathcal{F}$ . Then*

$$\mathbb{E}(\mathbb{E}(X | \mathcal{G}_2) | \mathcal{G}_1) = \mathbb{E}(X | \mathcal{G}_1)$$

*Proof.* (i) is a direct consequence of the linearity of the Lebesgue integral. For (ii) it suffices to notice that if  $G \in \mathcal{G}$ , the independence implies that

$$\int_G X \, d\mathbb{P} = \mathbb{E}(X \mathbb{1}_G) = \mathbb{E}X \mathbb{E} \mathbb{1}_G = \mathbb{E}X \mathbb{P}(G) = \int_G \mathbb{E}X \, d\mathbb{P}.$$

(iii) is a trivial consequence of the definition of conditional expectation and for (iv), we have to verify that  $\mathbb{E}(\mathbb{E}(X | \mathcal{G}_2) | \mathcal{G}_1)$  is  $\mathcal{G}_1$ -measurable and that

$$\int_{G_1} \mathbb{E}(\mathbb{E}(X | \mathcal{G}_2) | \mathcal{G}_1) d\mathbb{P} = \int_{G_1} X d\mathbb{P}$$

for all  $G_1 \in \mathcal{G}_1$ . By definition of the conditional expectation, the measurability is satisfied and the second identity follows as  $G_1 \in \mathcal{G}_1 \subseteq \mathcal{G}_2$  and we thus have

$$\int_{G_1} \mathbb{E}(\mathbb{E}(X | \mathcal{G}_2) | \mathcal{G}_1) d\mathbb{P} = \int_{G_1} \mathbb{E}(X | \mathcal{G}_2) = \int_{G_1} X d\mathbb{P}.$$

□

## A.2. Martingales and stopping times

**Definition A.2.1** (Martingale). *An stochastic process  $X$  that is adapted to a filtration  $(\mathcal{F}_t)_t$  with a finite expectation for all  $t$  is a martingale process when for almost surely for any  $s \leq t$ ,  $\mathbb{E}(X_t | \mathcal{F}_s) = X_s$ .*

An example of a martingale process is Brownian motion. The martingale property can be easily obtained from the definition of Brownian motion since for  $s < t$ , we have that  $\mathbb{E}(W_t | \mathcal{F}_s) = \mathbb{E}(W_t - W_s | \mathcal{F}_s) + \mathbb{E}(W_s | \mathcal{F}_s) = 0 + W_s$ .

**Definition A.2.2** (Stopping time). *A random variable  $\tau$  with respect to the filtration  $(\mathcal{F}_t)_t$  is a stopping time iff the event  $\{\tau \leq t\} \in \mathcal{F}_t$  for all  $t$ .*

**Definition A.2.3** (Local martingale). *An  $(\mathcal{F}_t)_t$ -adapted stochastic process  $X$  is an  $(\mathcal{F}_t)_t$ -local martingale if there exists a sequence  $\{\tau_k\}$  that almost surely increases to  $\infty$  such that the stopped process  $X^{\tau_k} = (X_{\tau_k \wedge t})_{t \geq 0}$  is an  $(\mathcal{F}_t)_t$ -martingale for all  $k$ .*

**Definition A.2.4** (semimartingale). *The process is called a semimartingale when it can be decomposed as  $X_t = M_t + A_t$  where  $M$  is a local martingale and  $A$  is a left-continuous process with existing limits to the right everywhere such that  $[A]_t < \infty$  for all  $t$ .*

Notice that from the definitions, it quickly follows that if a process is a martingale, that it is also a local martingale and a semimartingale.

## A.3. Quadratic (co)variation

In the previous section and in chapter 2.2, we already introduced the quadratic (co)variation. We can even go a step further than that.

**Definition A.3.1** (Variation of a semimartingale). *For a semimartingale  $X$  and any  $\alpha > 0$ , the  $\alpha$ -variation is defined as*

$$V_t^\alpha(X) = \lim_{\|\Delta\| \rightarrow 0} \sum_{i=1}^N \|X_{t_i} - X_{t_{i-1}}\|^\alpha$$

where  $\Delta = \{0 = t_0 < t_1 \dots < t_N = t\}$  is any partition of  $[0, t]$  and  $\|\Delta\| = \sup |t_i - t_{i-1}|$ .

**Definition A.3.2** (Total and quadratic variation). *If  $\alpha = 1$ , we refer to  $V_t^\alpha$  as the total variation and if  $\alpha = 2$  we refer to  $V_t^\alpha$  as the quadratic variation. Note that this coincides with our definition of quadratic covariation given earlier.*

The next theorem demonstrates that there is a difference between finite total variation and finite quadratic variation in practice.

**Theorem A.3.3.** *Let  $X$  be a continuous semimartingale. If there is an  $\alpha > 0$  such that  $0 < V_t^\alpha(X) < \infty$  then  $V_t^\gamma(X) = 0$  for all  $\gamma > \alpha$  and  $V_t^\beta(X) = \infty$  for all  $0 < \beta < \alpha$ .*

*Proof.* Since  $X$  is continuous, we can choose our partition  $\Delta$  such that  $\|X_{t_i} - X_{t_{i-1}}\| < 1$  for all  $i$ . Then we must have that

$$\sum_{i=1}^N \|X_{t_i} - X_{t_{i-1}}\|^\gamma \leq \sum_{i=1}^N \|X_{t_i} - X_{t_{i-1}}\|^\alpha \leq \sum_{i=1}^N \|X_{t_i} - X_{t_{i-1}}\|^\beta$$

Hence, when  $\|\Delta\| \rightarrow 0$ , we obtain  $V_t^\gamma(X) \leq V_t^\alpha(X) \leq V_t^\beta(X)$ . Now since we choose  $\alpha > \gamma$ , we must also have that

$$\sum_{i=1}^N \|X_{t_i} - X_{t_{i-1}}\|^\gamma \leq \sum_{i=1}^N \|X_{t_i} - X_{t_{i-1}}\|^\alpha \max_i \|X_{t_i} - X_{t_{i-1}}\|^{\gamma-\alpha}$$

When  $\|\Delta\| \rightarrow 0$ , the sum remains finite as  $V_t^\alpha(X) < \infty$  while the maximum tends to 0 by continuity of  $X$  and hence  $V_t^\gamma(X) = 0$ . We can perform a similar trick for  $\beta$ , since  $\beta < \alpha$ , we have that

$$\sum_{i=1}^N \|X_{t_i} - X_{t_{i-1}}\|^\alpha \leq \sum_{i=1}^N \|X_{t_i} - X_{t_{i-1}}\|^\beta \max_i \|X_{t_i} - X_{t_{i-1}}\|^{\alpha-\beta}$$

By continuity of  $X$  we now see that  $V_t^\alpha(X) > 0$  can only be true when  $V_t^\beta(X) = \infty$  □

**Corollary A.3.4.** *For a Brownian motion  $W$ , we have proven that  $V_t^2(W) = t$  and therefore  $W$  is not of finite total variation.*

**Theorem A.3.5.** *A semimartingale  $X$  in  $\mathbb{R}$  is of finite total variation if and only if it is a linear combination of monotone processes.*

*Proof.* Note that for a monotone process  $Y$  in  $\mathbb{R}$ , the total variation forms a telescopic sequence and hence  $V_t^1(Y) = |Y_t - Y_0| < \infty$ . From the triangle inequality, it also follows that any linear combination of processes of finite total variation is of finite total variation. For the converse, notice that  $X_t = V_t^1(X) - (V_t^1(X) - X_t)$ . Since  $t \mapsto V_t^1(X)$  is clearly a monotone function, we are done when we show that  $t \mapsto V_t^1(X) - X_t$  is monotone. For this, we point out that the triangle inequality implies that for  $s < t$ , we have that  $|X_t - X_s|$  will be below the total variation between  $s$  and  $t$  and therefore  $X_t - X_s \leq V_t^1(X) - V_s^1(X)$ . Hence  $t \mapsto V_t^1(X) - X_t$  is an increasing function. □

**Corollary A.3.6.** *The quadratic covariation of two semimartingales  $X$  and  $Y$  is of finite total variation.*

*Proof.* Recall from proposition 2.2.6 that we can write the quadratic covariation as

$$[X, Y]_t = \frac{[X + Y]_t - [X]_t - [Y]_t}{2}$$

From the definition, it should be clear that the quadratic variation is an increasing process and thus the quadratic covariation is a linear combination of increasing process and therefore of finite variation. □

**Corollary A.3.7.** *Suppose that  $X$  and  $Y$  are continuous semimartingales and one is of finite total variation. Then  $[X, Y] = 0$ .*

*Proof.* Since  $(X, Y) \mapsto [X, Y]$  defines a bilinear map, we can apply the Cauchy-Schwarz inequality to find that

$$\|[X, Y]\| \leq \sqrt{\|[X]\| \|[Y]\|}$$

If either  $X$  or  $Y$  is of finite total variation, the quadratic variation equals 0 and therefore the right hand side equals 0, thus proving the desired result. □

**Theorem A.3.8.** *Let  $X$  and  $Y$  be local martingales. Then there exists a process  $[X, Y]$  of locally finite variation that is almost surely unique such that  $[X, Y]_0 = 0$  and  $XY - [X, Y]$  is a local martingale.*

*Proof.* see Kallenberg, 2006 [10]. □

As the notation in theorem A.3.8 suggests, the process that converts  $XY$  into a local martingale is indeed the quadratic covariation of  $X$  and  $Y$ , see Karandikar and Rao, 2014 [11]. From this uniqueness, one can also deduce that the quadratic covariance indeed defines a bilinear map.

**Corollary A.3.9.** *For a standard Brownian motion  $W$ , we have  $[W] = t$ .*

*Proof.* By theorem A.3.8, it suffices to show that the process  $W_t^2 - t$  is a martingale. From a direct calculation for  $s < t$ , we obtain that

$$W_t^2 - t = (W_s + W_t - W_s)^2 - t = W_s^2 + 2(W_t - W_s)W_s + (W_t - W_s)^2 - t$$

We thus derive that

$$\begin{aligned}\mathbb{E}(W_t^2 - t \mid \mathcal{F}_s) &= \mathbb{E}(W_s^2 \mid \mathcal{F}_s) + 2\mathbb{E}(W_s(W_t - W_s) \mid \mathcal{F}_s) + \mathbb{E}((W_t - W_s)^2 \mid \mathcal{F}_s) - t \\ &= W_s^2 + 2W_s\mathbb{E}(W_t - W_s) + \mathbb{E}((W_t - W_s)^2) - t \\ &= W_s^2 + t - s - t = W_s^2 - s\end{aligned}$$

Hence  $W_t^2 - t$  is indeed a martingale and thus  $[W]_t = t$ . □

# B

## Markov Chain Monte Carlo methods and the Metropolis-Hastings algorithm

In statistics, we encounter many situation where a distribution is intractable. For example, in this thesis we have seen that the transition density of a stochastic process is usually intractable, leading to an intractable distribution of the center of data. Markov Chain Monte Carlo methods form a collection of techniques that can be used to draw samples from such distributions.

### B.1. Discrete time Markov chains

Let  $\{X_n\}$  be a sequence of random variables in a measurable state space  $(\mathcal{X}, \mathcal{B})$  with an underlying probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . Furthermore let  $\mathcal{F}_n^X = \sigma(X_k : k \leq n)$  denote the sigma-field generated by  $X_0, \dots, X_n$ . We recall the definition of a transition kernel

**Definition B.1.1** (Transition kernel). *A transition kernel on  $\mathcal{X}$  is a map  $P : \mathcal{X} \times \mathcal{B} \rightarrow [0, 1]$  such that*

- (i) *For every  $x \in \mathcal{X}$ , the map  $B \mapsto P(x, B)$  defines a probability measure on  $(\mathcal{X}, \mathcal{B})$ .*
- (ii) *For every  $B \in \mathcal{B}$ , the map  $x \mapsto P(x, B)$  is a measurable map  $(\mathcal{X}, \mathcal{B}) \rightarrow ([0, 1], \mathcal{B}[0, 1])$ .*

*If  $P(x, \cdot)$  admits a density  $p$ , this density is referred to as the transition density and satisfies*

$$P(x, B) = \int_B p(x, y) \, \mathrm{d}y, \quad B \in \mathcal{B}$$

**Definition B.1.2.**  *$X = \{X_n\}$  is a Markov chain with initial distribution  $\nu$  if  $\mathbb{P}(X_0 \in B) = \nu(B)$  for all  $B \in \mathcal{B}$  and if there exists transition kernel  $P$  such that the Markov property holds:*

$$\mathbb{P}(X_{n+1} \in B \mid \mathcal{F}_n^X) = \mathbb{P}(X_{n+1} \in B \mid \sigma(X_n)) = P(X_n, B)$$

**Remark B.1.3.** *If  $\mathcal{X}$  is a countable space, we conveniently write  $P(x, y)$  for  $P(x, \{y\})$ .*

We can define  $P^m$  inductively by

$$P^m(x, B) = \int_{\mathcal{X}} P(y, B) P^{m-1}(x, \mathrm{d}y)$$

and for measurable functions, we can define

$$Pf(x) = \int_{\mathcal{X}} f(y) P(x, \mathrm{d}y)$$

The Markov property then yields

$$\mathbb{E}(\mathbb{1}_B(X_{n+1}) \mid \mathcal{F}_n^X) = P\mathbb{1}_B(X_n), \quad B \in \mathcal{B}.$$

Since  $P$  and  $\mathbb{E}$  are linear operators and measurable function can be approximated by simple functions, we deduce that

$$\mathbb{E}(f(X_{n+1}) | \mathcal{F}_n^X) = Pf(X_n)$$

for any measurable  $f : \mathcal{X} \rightarrow \mathbb{R}$ . This property implies that

$$\mathbb{P}(X_{n+2} \in B | \mathcal{F}_n^X) = \mathbb{E}(\mathbb{E}(\mathbb{1}_B(X_{n+2}) | \mathcal{F}_{n+1}^X) | \mathcal{F}_n^X) = \mathbb{E}(P\mathbb{1}_B(X_{n+1}) | \mathcal{F}_n^X) = P(P\mathbb{1}_B)(X_n) = P^2(x, B)$$

In general, it can be deduced that  $\mathbb{P}(X_{n+m} \in B | \mathcal{F}_n^X) = P^m(x, B)$ .

**Definition B.1.4.** An invariant distribution for  $X$  on  $\mathcal{X}$  is a probability distribution  $\Pi$  on  $\mathcal{X}$  such that

$$\Pi(B) = \int_{\mathcal{X}} P(x, B) \Pi(dx), \quad B \in \mathcal{B}$$

For a Markov chain  $X$  with invariant distribution  $\pi$ , it can be shown (see Athreya et al., 1996 [2]) that under certain conditions,

$$\sup_{B \in \mathcal{B}} |P^n(x, B) - \Pi(B)| \rightarrow 0, \quad n \rightarrow \infty$$

for almost all  $x \in \mathcal{X}$ . In this case,  $\Pi$  is called the *stationary distribution* of  $X$ . If a Markov chain has a unique stationary distribution, the Markov chain is called *ergodic*.

## B.2. Markov chain Monte Carlo methods

**Definition B.2.1** (Markov chain Monte Carlo). A Markov chain Monte Carlo (MCMC) method for the simulation of a distribution  $\Pi$  is any method that returns an ergodic Markov chain with stationary distribution  $\Pi$ .

The most well-known MCMC method is the Metropolis-Hastings (MH) algorithm.

**Definition B.2.2** (Metropolis-Hastings (MH) algorithm). Suppose that  $\Pi$  has density  $\pi$ . The MH algorithm constructs a Markov chain by updating  $x_n = x$  to  $x_{n+1}$  in the following steps.

1. Propose  $y$  from a proposal density  $q(x, \cdot)$
2. Compute

$$\alpha(x, y) = \min \left\{ 1, \frac{\pi(y) q(y, x)}{\pi(x) q(x, y)} \right\}$$

3. Accept the proposal  $y$  with probability  $\alpha(x, y)$  and set  $x_{n+1} = x$  or reject  $y$  with probability  $1 - \alpha(x, y)$  and set  $x_{n+1} = x$ .

Note that there is a lot of freedom in the choice of  $q$ . If  $q$  is chosen symmetrically in  $x$  and  $y$ , we observe that the MH algorithm accepts a proposal  $y$  if  $y$  is more likely than  $x$  according to  $\pi$  and has a probability of rejecting if this is not the case. It is a well-known result that the Markov chain constructed in B.2.2 indeed has  $\pi$  as invariant distribution.

The MH algorithm is in particular greatly appreciated by Bayesian statisticians as the algorithm only needs an expression for  $\pi$  up to a proportionality constant.

### B.2.1. Langevin adjusted proposals

Langevin adjusted proposals are proposals done with the aim to obtain faster convergence of the MH algorithm to the stationary distribution. In a similar way as Roberts and Rosenthal, 1998 [19], we implement an accept/reject step in the Euler discretization of the Langevin diffusion  $(\Lambda_t)_t$  that solves the stochastic differential equation given by

$$d\Lambda_t = \frac{h}{2} \nabla \ell(\Lambda_t) dt + \sqrt{h} dW_t$$

Here,  $\Lambda_t$  can be interpreted as a given location and the SDE describes a stochastic movement towards high values of  $\ell$ , as indicated by the  $\nabla \ell$  term in the drift. Under weak conditions, this SDE has a unique

solution that is a continuous-time Markov process whose asymptotic stationary distribution is  $\exp \ell$ , see Michelot et al., 2018 [16].

The Euler discretization of the Langevin SDE is given by

$$\Lambda_{t+1} = \Lambda_t + \frac{h}{2} \nabla \ell(\Lambda_t) + \sqrt{h} Z, \quad Z \sim \mathcal{N}(0, I) \quad (\text{B.1})$$

Now observe that if we implement this into a Metropolis-Hastings algorithm, we are drawing our proposed updates  $\theta^\circ$  when we are at  $\theta$  from an  $\mathcal{N}(\theta + \frac{h}{2} \nabla \ell(\theta), hI)$ -distribution. Hence the proposal density  $q$  is given by, the probability density function of a  $\mathcal{N}(x + \frac{h}{2} \nabla \ell(x), hI)$ -distributed random variable evaluated in  $y$ , i.e.

$$q(x, y) = (2\pi h)^{-\frac{3}{2}} \exp \left\{ -\frac{1}{2h} \left( y - x - \frac{h}{2} \nabla \ell(x) \right)^T \left( y - x - \frac{h}{2} \nabla \ell(x) \right) \right\} \quad (\text{B.2})$$



# Index

- $k$ -jets, 39
- Atlas, 3
- Brownian motion
  - In  $\mathbb{R}$ , 15
  - In  $\mathbb{R}^d$ , 16
  - Infinitesimal generator, 23
- Chart, 3
- Christoffel symbols, 8
- Connection, 8
- Covariant derivative, 10
- Diffusion bridge, 41
- Diffusion process, 18
  - Infinitesimal generator, 23
- Directional derivative, 4
- Divergence operator, 12
- Einstein notation, 7
- Euler-Maruyama method, 35
- Exponential map, 10
- Fokker-Planck equation, 24
- Geodesic, 10
  - Geodesic equations, 10
- Geodesic distance, 10
- Global section, 6
- Gradient operator, 12
- Infinitesimal Generator, 22
- Itô integral, 17
- Itô's lemma, 19
- Kolmogorov equations, 23
- Langevin updated proposals, 47
- Laplace-Beltrami operator, 13
- Levi-Civita connection, 8
  - Christoffel symbols of the Levi-Civita connection, 8
- Lie bracket, 6
- Local coordinate systems, 3
- Logarithmic map, 10
- Manifold, 3
  - Differentiable manifold, 3
  - Smooth manifold, 3
- Markov chain, 65
- Markov Chain Monte Carlo, 46
- Markov process, 21
- Martingale, 62
  - Local martingale, 62
  - Semimartingale, 62
- Parabolic equation, 24
  - Existence of a semigroup of weak solutions, 25
  - Weak solution, 25
- Projection onto tangent space, 31
- Regular submanifold of  $\mathbb{R}^N$ , 29
  - Brownian motion, 31
  - Riemannian metric, 30
- Riemannian manifold, 7
  - Riemannian metric, 7
- Stationary distribution, 66
- Stochastic gradient descent, 51
  - Exponential map, 51
  - Normalizing updates, 52
- Stochastic process, 15
  - Adapted to a filtration, 15
  - Continuity, 15
  - Independence of stochastic processes, 15
  - Law, 15
  - Quadratic (co)variation, 16
  - Sample path, 15
- Stopping time, 62
- Stratonovich integral, 19
- Tangent bundle, 4
- Tangent space, 4
- The circle, 7
  - Brownian motion in local coordinates, 28
  - Exponential map, 11
  - Laplace-Beltrami operator, 14
  - Riemannian metric, 7
  - Tangent space, 7
- The sphere, 7
  - Brownian motion in local coordinates, 28
  - Christoffel symbols of the Levi-Civita connection on the Sphere, 9
  - Exponential map on the sphere, 11
  - Geodesics on the sphere, 11
  - Laplace-Beltrami operator, 14
- The torus
  - Brownian motion, 29

- Christoffel symbols of the Levi-Civita connection, [29](#)
- Riemannian metric, [28](#)
- Torus, [28](#)
- Transition density, [22](#)
- Transition kernel, [21](#)
  
- Usual conditions, [15](#)
  
- Vector bundle, [5](#)
- Vector field, [6](#)
  - Vector fields along a curve, [9](#)