Contributions to Stein's Method on Riemannian Manifolds

Alexander Lewis

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

February 6, 2023

Abstract

The overarching theme of this thesis is the study of Stein's method on manifolds. We detail an adaptation of the density method on intervals in $\mathbb{R}$ to the unit circle $\mathbb{S}^1$ and give examples of bounds between circular probability distributions. We also use a recently proposed framework to bound the Wasserstein distance between a number of probability measures on Riemannian manifolds with both positive and negative curvature. Particularly, a finite parameter bound on the Wasserstein metric is given between the Riemannian-Gaussian distribution and heat kernel on $\mathbb{H}^3$, which gives a finite sample bound of the Varadhan asymptotic relation in this instance. We then develop a new framework to extend Stein's method to probability measures on manifolds with a boundary. This is done by the addition of a local time term in the diffusion. We find that many results carry over, with appropriate modifications, from the boundary-less case.

Acknowledgements

I would like to give a tremendous thank you to my supervisors Christopher Fallaize and Karthik Bharath who have given me so much support over the last few years. An even bigger thank you to Huiling Le. She is the best supervisor anyone could ask for, and I wish her the happiest of retirements.

To the many people of the Mathematical Sciences department here at Nottingham that have provided me with tremendous amounts of support for the last 4 years: Akash Sharma, Charles Valdez, Celene Lee, Valentin Breaz, Tom Laird, Anne Boschman, Adam Blakey, Emily Mitchell, Harry Wells, Tom Radley, Wasiur KhudaBukhsh, Michael Tretyakov, Theodore Kypraios and Cameron Bunney.

To my friends, both online and offline: Matthew Thompson, Marc Warner, Bartosz Kozierkievicz, Claire Burchfield, Răzvan Alexe, Bastien Gelly, Abdulrahman AlHomoud, Kieran Martin, Avin Ekanayake, Ronald Vuong, Trevor Pham, Sarah Kaye and George Stockhall.

To my parents and to my late grandmother, Avarina Humphries.

This thesis is dedicated to the Strapping Young Lad album ‘City’.

“We can’t expect God to do all the work.”

Joshua Graham

Fallout: New Vegas

Contents

Chapter 1

Introduction

1.1 Stein's Method

It was in 1972 that Charles Stein first published [Ste72], the foundations to the method that would later bear his name. He originally used this method as a tool for providing an alternative proof for the central limit theorem for sums of random variables. This work was later refined and formalised in [Ste86] which can also be recognised as the real starting point for Stein's method.

The main objective of Stein's method is to find a way to bound an integral probability metric between two random variables X, Y

$$d_{\mathcal{H}}(X, Y) = \sup_{h \in \mathcal{H}} |\mathbb{E}[h(X)] - \mathbb{E}[h(Y)]| \tag{1.1}$$

where $\mathcal{H}$ is some space of test functions.

Stein's idea, in principle, is simple. He found a characterization of the normal distribution in terms of a differential operator. With that differential operator in hand, it was possible for him to bound integral probability metrics, specifically the Kolmogorov metric, and show convergence in distribution between the normal distribution and sums of non-independent random variables.

The characterization can be expressed within the eponymous lemma:

Lemma 1.1.1 (Stein's Lemma). *Suppose* $Z \sim \mathrm{N}(\mu, \sigma^2)$ *and let* $g \in C^1(\mathbb{R})$ *be a function such that* $(\mathrm{Id} - \mu)g \in L^1(Z)$ *and* $g' \in L^1(Z)$*, then*

$$\mathbb{E}[(Z - \mu)g(Z)] = \sigma^2 \mathbb{E}[g'(Z)].$$

The operator $\mathcal{A}f(x) = \sigma^2 f'(x) - (x - \mu)f(x)$ is known as the canonical Stein operator for the normal distribution, and Stein's lemma can be rewritten in terms of the operator so that $\mathbb{E}[\mathcal{A}f(Z)] = 0$.

The ingenuity of the method is presented to us via the Stein equation. Stein decided to formulate an auxiliary function by equating the Stein operator and a function that resembles the formulation of the metric;

$$\sigma^2 f_h'(x) - (x - \mu)f_h(x) = h(x) - \mathbb{E}[h(Z)]. \tag{1.2}$$

The function f_h is known as the solution to the Stein equation. It is clear that when we take expectations under Z on the right hand side, we obtain 0. The left hand side also agrees with this outcome by recognising Stein's lemma.

The metric can be obtained by taking expectations with respect to another random variable X,

$$d_{\mathcal{H}}(X, Z) = \sup_{h \in \mathcal{H}} |\mathbb{E}[h(X)] - \mathbb{E}[h(Z)]| = \sup_{h \in \mathcal{H}} |\mathbb{E}[\sigma^2 f_h'(X) - (X - \mu)f_h(X)]|. \tag{1.3}$$

This single equation permits us to reduce the original problem of calculating (1.1), a rather difficult one involving two random variables, into a more manageable problem involving a single random variable. By bounding the right hand side of (1.3), we can bound the metric without having to compute any expectations with respect to Z. To do this, we typically need bounds on the solution to the Stein equation f_h. The Stein equation can be explicitly solved, in the case of a standard normal distribution. It takes the form

$$f_h(x) = e^{x^2/2} \int_{-\infty}^{x} (h(u) - \mathbb{E}[h(Z)])e^{-u^2/2} du. \tag{1.4}$$

We previously mentioned that $\mathcal{H}$ was some set of test functions, but we did not elaborate about what kind of test functions are typically taken. There are three different metrics which are primarily used when comparing distributions (each metric is associated with a unique class of test functions), which are the following:

1. The Kolmogorov Metric d_K — here we take $\mathcal{H} = \{\mathbb{I}_{\{\cdot \leq z\}} : z \in \mathbb{R}\}$ which describes the maximum distance between CDFs.

2. The Total Variation Metric d_{TV} — here $\mathcal{H} = \{\mathbb{I}_{\{\cdot \in A\}} : A \in \mathcal{B}(\mathbb{R}^n)\}$, a generalization of the Kolmogorov metric.

3. The Wasserstein metric d_W — here we take $\mathcal{H} = \{h : |h(x) - h(y)| \leq |x - y|, x, y \in \mathbb{R}^n\}$, which is also known as the earth moving distance, is the metric that is central to optimal transport.

Depending on what set of test functions one takes the supremum over, the behaviour of the solution (1.4) will change. For example, in both Kolmogorov and Total Variation cases, every test function is not differentiable (in fact not even continuous) and so we are restricted with how many derivatives we can take of f_h.

This work will be primarily concerned with the bounding of the Wasserstein metric and not of any other. There are two reasons for this: First, calculations for the framework that we will introduce in Section 1.1.3 require test functions to be differentiable. Second, convergence in the Wasserstein metric implies weak convergence. More concretely, suppose X is a random variable and $\{X_n\}_{n\in\mathcal{I}}$ are a sequence of random variables. If $\lim_{n\to\infty} d_W(X_n, X) = 0$, then $X_n \xrightarrow{P} X$ as $n \to \infty$ where $\xrightarrow{P}$ is convergence in probability.

Cumulative distribution functions do not make sense to have on a manifold, and so the Kolmogorov metric does not exist. Bounds on the total variation metric have been calculated on $\sqrt{n}\mathbb{S}^n$ (the n-sphere of radius $\sqrt{n}$) in [Mec09] and [DF87], but both use mainly geometric approaches.

To bound the solution, one can utilise properties of the CDF of the normal distribution. Below is a standard result within Stein's method on the properties of (1.4) [CGS10]:

Lemma 1.1.2. *For a given function $h : \mathbb{R} \to \mathbb{R}$, let f_h be the solution* (1.4) *to the Stein equation* (1.2). *If h is bounded, then*

$$\|f_h\|_\infty \leq \sqrt{\frac{\pi}{2}} \|h - \mathbb{E}[h(Z)]\|_\infty \quad \text{and} \quad \|f_h'\|_\infty \leq 2 \|h - \mathbb{E}[h(Z)]\|_\infty .$$

If h is absolutely continuous, then

$$\|f_h\|_\infty = \|f_h''\|_\infty = 2 \|h'\|_\infty , \quad \text{and} \quad \|f_h'\|_\infty \leq \sqrt{\frac{2}{\pi}} \|h'\|_\infty .$$

Here, the notation $\|\cdot\|_\infty$ is the infinity norm of a function, i.e. $\|f\|_\infty = \sup_{x\in\mathbb{R}^n} |f(x)|$. The first half of this lemma is used to bound the Kolmogorov metric, since $h_z(x) = \mathbb{I}_{x\leq z}$ is not continuous at $x = z$. The latter half is used for bounding the Wasserstein metric, since it is the set of functions with Lipschitz constant 1, so these inequalities simplify on applying $\|h'\|_\infty = 1$.

With the solution and its derivatives bounded we may use any tools in the literature to bound $\mathbb{E}[\mathcal{A}f_h(X)]$ given the above lemma. This includes, but is not limited to: sums of random variables, exchangeable pairs, a-stein pairs, size-bias and zero-bias coupling — see [Ros11] for a brief overview of all of these topics. The choice one makes to use one of the above is usually based upon the context of the problem at hand.

Not only is a Stein method readily available for the normal distribution, but also for many popular univariate distributions such as: Gamma [Luk94], Beta [Döb15], Exponential [CFR11] and Laplace [PR12] distributions. This is not restricted to just continuous distributions, but to also discrete ones such as: Poisson [Che75], Geometric [Pek96] Binomial [LRS17b], Negative Binomial [BP99], Multinomial [Loh92] distributions. Our focus, however, shall be on absolutely continuous distributions as the methods to construct Stein operators and the Stein equation are able to encapsulate a useful subset of the space of probability measures.

Applications of Stein's method are now very abundant and stretch from proving central limit theorems to applying methods in machine learning algorithms, some of which are discussed below. One particularly popular use of Stein's method is known as the Kernel Stein Discrepancy (KSD) which is a modification of the original objective discrepancy (1.3) which extends it to more test spaces and distributions;

$$S(p, q, \mathcal{G}) = \sup_{g \in \mathcal{G}} |\mathbb{E}_q[\mathcal{A}_p g(Y)]|$$

where $X \sim p$, $Y \sim q$ and $\mathcal{A}_p$ is a kernelized Stein operator for X. As an example, a well used test space is the unit ball $\mathcal{H} = \{g \in \mathcal{G} : \|g\| = 1\}$. Though the KSD may not generate a metric space, it does capture important qualities of distributional convergence [GM15]. The KSD was introduced simultaneously in [LLJ16] and [CSG16] (both appearing in ICML 2016) to be used as a tool in goodness-of-fit testing. The concept is to construct a reproducing kernel Hilbert space $\mathcal{H}$ out of $\mathcal{G}$ so that a general function $f \in \mathcal{H}$ may be written as $f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}}$ for some kernel k by the reproducing property.

The KSD is a robust framework that has been utilised in many areas of computational statistics. Applications include: measuring sample quality of Monte Carlo variates [GM15]; a modification of the gradient descent known as Stein variational gradient descent [LW16]; optimally thinning an MCMC output [RCC$^+$20]; numerically approximating integrals via Monte Carlo methods [BOPG18].

Though it is clear the research into statistical uses of Stein's method has been fruitful, probabilistic theory has also made strides. Nourdin and Peccatti established a connection between Stein's method and Malliavin calculus in [NP09] which later led on to multivariate normal approximations [NPR10]. See [NP12] for a general overview on the matter. The Malliavin calculus approach is regarded as the third approach to Stein's method; the other two, the classical density and diffusion approaches, are the main topics of focus within this work.

Before going into detail, we briefly present the density and diffusion approaches for a general absolutely continuous distribution with density function $p(x)$ in the multidimensional setting. We then finalize the introductory part of Stein's method with a short study on the extension to manifold.

1.1.1 Density Approach

Suppose X is a probability distribution with density function $p(x)$ on any open or closed variant of the interval $[a, b]$, $a < b$. We first want to find a Stein operator, and to do that we look back on the form of the canonical Stein operator in Lemma 1.1.1 and make the ansatz that the operator $\mathcal{A}$ is a first order differential operator

$$\mathcal{A}f(x) = f'(x) + Ap(x)f(x),$$

where A is some functional acting on p. To find what A looks like, we calculate $\mathbb{E}[f'(X)]$. By integration by parts,

$$\begin{aligned}\mathbb{E}[f'(X)] &= \int_a^b f'(x)p(x)dx,\\ &= f(x)p(x)\Big|_a^b - \int_a^b f(x)p'(x)dx.\end{aligned}$$

Assuming that $f(x)p(x)|_a^b = 0$, we are left with

$$\mathbb{E}[f'(X)] + \int_a^b f(x)p'(x)dx = 0$$

This expression is 0, and since the expectation of the Stein operator must be 0 under p, the integral on the right hand side must relate to A somehow. We can make the integral an expectation by writing

$$\int_a^b f(x)p'(x)dx = \int_a^b f(x)\frac{p'(x)}{p(x)}p(x)dx = \mathbb{E}[f(X)(\log p)'(X)].$$

We therefore have the more general form of the Stein lemma:

Lemma 1.1.3. *Suppose X is a probability distribution with density p on an interval $-\infty \leq a < b \leq \infty$ which satisfies $p' \in L^1(dx)$. Moreover, let $\mathcal{F}(X)$ be some class of functions, depending on X, satisfying $f' \in L^1(X)$ and $f(a+)p(a+) = f(b-)p(b-)$. Then*

$$\mathbb{E}\left[f'(X) + f(X)\frac{p'(X)}{p(X)}\right] = 0.$$

We may also write the Stein operator in the more compact form

$$\mathcal{A}_p f = \frac{(fp)'}{p}.$$

Remark. We note that there is no one definitive Stein operator. In fact there are infinitely many Stein operators one could choose, for example in the normal case, we could either choose from

$$\mathcal{A}_1 f(x) = \sigma^2 f'(x) - (x-\mu)f(x)$$

directly from the Stein lemma 1.1.1, or

$$\mathcal{A}_2 f(x) = f'(x) - \frac{(x-\mu)}{\sigma^2}f(x)$$

or even

$$\mathcal{A}_3 f(x) = \sigma^2 f''(x) - (x-\mu)f'(x)$$

by using the lemma for general densities above. All three will give $\mathbb{E}[\mathcal{A}_i f(X)] = 0$ meaning that either are valid Stein operators.

Henceforth, after obtaining a way to generate a Stein operator for a given density p, the procedure detailed at the start of this chapter is followed: constructing the Stein equation, and bounding the solution and its derivatives. For a detailed exposition of the general density method for densities on $\mathbb{R}$, we refer the reader to [LRS17b].

A multidimensional extension of this general density method is presented in [MRS18] in which one can also choose the type of derivative in the Stein operator. For example, one can use the gradient function to define a vectorised Stein operator

$$\mathcal{T}_p f = \frac{\nabla(fp)}{p}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, or a divergence type operator

$$\mathcal{T}_p f = \frac{\nabla \cdot (fp)}{p}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ is vectorised. The procedure of standardization of the Stein operator is detailed in [MRS18, Section 3]. Briefly, these are groups of operators that can be transformed interchangeably despite having different supports. For example, the operators $\mathcal{A}_1 f(x) = f'(x) - xf(x)$ and $\mathcal{A}_2 g(x) = g''(x) - xg'(x)$ for the standard normal distribution are essentially the same operator. General solutions to the multidimensional version of the Stein equation are not readily available for general densities p. In an effort to move towards generality, we now detail a different approach to Stein's method.

1.1.2 Diffusion Approach

Barbour was the first to notice [Bar90] that the infinitesimal generator of the Ornstein-Uhlenbeck (OU) process can be used as a Stein operator for the normal distribution. It is also true that the normal distribution is the stationary distribution of the OU process, and therefore its invariant distribution. The Ornstein-Uhlenbeck process is the solution to the Stochastic Differential Equation (SDE)

$$dX_t = dB_t - \frac{1}{2} X_t dt$$

where $\{B_t\}_{t \in \mathbb{R}^+}$ is a standard Brownian motion on $\mathbb{R}$. Define the semigroup of $\{X_t\}_{t \in \mathbb{R}}$, $P_t f(x) = \mathbb{E}[f(X_t) | X_0 = x]$. Then the infinitesimal generator of the

Ornstein-Uhlenbeck process is

$$\mathcal{A}f(x) := \lim_{t\to 0}\frac{P_t f(x) - f(x)}{t} = \frac{1}{2}f''(x) - \frac{1}{2}xf'(x)$$

which, after making the substitution $f'(x) = g(x)$, equals the Stein operator in Lemma 1.1.1.

With this new operator in hand, we can rewrite the Stein equation,

$$\frac{1}{2}f''(x) - \frac{1}{2}xf'(x) = h(x) - \mathbb{E}[h(X)].$$

Despite the Stein operators of the diffusion and density approaches being similar, the solution (when put in the context of the diffusion approach) is vastly different when compared with the solution (1.4);

$$f_h(x) = \int_0^\infty \mathbb{E}[h(X)] - P_t h(x) dt. \tag{1.5}$$

It is straightforward to generalise this approach for distributions with densities p on support $\mathbb{R}^n$ by borrowing the notion of the over-damped Langevin diffusion from statistical physics. This is an extension of the OU process that is the solution to the SDE

$$dX_t = dB_t - \frac{1}{2}\nabla(\log p)(X_t)dt \tag{1.6}$$

where $\{B_t\}_{t\in\mathbb{R}^+}$ is now a standard Brownian motion on $\mathbb{R}^n$. The infinitesimal generator is

$$\mathcal{A}f(x) = \frac{1}{2}\Delta f(x) + \frac{1}{2}\langle\nabla \log p(x), \nabla f(x)\rangle,$$

for a function $f \in C^2(\mathbb{R}^n)$ and where $\langle\cdot,\cdot\rangle$ is the inner product on $\mathbb{R}^n$. To simplify, we assume that the form of p is $p \propto e^{-\phi}$ where $\phi \in C^2(\mathbb{R})$ is non-negative, giving

$$\mathcal{A}f(x) = \frac{1}{2}\Delta f(x) - \frac{1}{2}\langle\nabla\phi(x), \nabla f(x)\rangle. \tag{1.7}$$

The Stein equation is now a partial differential equation (PDE)

$$\frac{1}{2}\Delta f_h(x) - \frac{1}{2}\langle \nabla\phi(x), \nabla f_h(x)\rangle = h(x) - \mathbb{E}[h(X)] \tag{1.8}$$

and unsurprisingly, its solution is identical to the solution in the normal case on $\mathbb{R}$ (1.5)

$$f_h(x) = \int_0^\infty \mathbb{E}[h(X)] - P_t h(x) dt. \tag{1.9}$$

Remark. The PDE (1.8) is known as the Poisson equation (or weighted Poisson equation to some) in the PDE theory, more commonly written as $Lf = \bar{h}$ where L is a second order elliptic differential operator and $\mathbb{E}[\bar{h}] = 0$. The probabilistic representation of the solution (1.9) pre-dates Barbour's paper on the diffusion approach. See [Fre16] for the original derivation.

Bounds on the solution (1.9) are presented in [MG16] and can be used in conjunction with the Stein equation to bound the Wasserstein metric. To obtain these, an important assumption on the behaviour of the density p is required, namely, p is strongly log concave.

Definition 1.1.4. A function $f \in C^2(\mathbb{R}^n)$ is κ-strongly concave if

$$\mathrm{Hess}^f(v,v)(x) \leq -\kappa|v|^2$$

for $\kappa > 0$ for all $x, v \in \mathbb{R}^n$.

The notation $\mathrm{Hess}^f(v,v)(x) = v^\intercal \mathrm{Hess}(f(x))v$ is useful to have when one wants to extend notions to manifold.

If one assumes the particular form $p \propto e^{-\phi}$ of p for some function ϕ, the strong log concave assumption on p simplifies to

$$\mathrm{Hess}^\phi(v,v)(x) \geq \kappa|v|^2.$$

We then have the following theorem [MG16]:

Theorem 1.1.5. *Suppose that* $\log p \in C^4(\mathbb{R}^n)$ *is* κ*-strongly concave with* $C_2(\log p) \leq L_2$ *and* $C_3(\log p) \leq L_3$, $L_2, L_3 > 0$. *For each* $x \in \mathbb{R}^n$, *let* $\{X_t\}_{t\in\mathbb{R}^+}$ *represent the overdamped Langevin diffusion with infinitesimal generator* (1.7) *and initial position* $X_0 = x$. *Then, for each Lipschitz* $h \in C^3(\mathbb{R}^n)$, *the solution* (1.9) *solves the Stein equation* (1.8) *and satisfies*

$$C_0(f_h) \leq \frac{2}{\kappa}C_0(h), \quad C_1(f_h) \leq \frac{2L_2}{\kappa^2}C_0(h) + \frac{1}{\kappa}C_1(h),$$

$$C_2(f_h) \leq \left(\frac{6L_2^2}{\kappa^3} + \frac{L_3}{\kappa^2}\right)C_0(h) + \frac{3L_2}{\kappa^2}C_1(h) + \frac{2}{3\kappa}C_2(h).$$

Here we have labelled $C_i(f)$ to be the Lipschitz constant of the ith derivative of the function f.

1.1.3 Extension to Manifold

The final generalization we make in this introduction will be to extend the notion of the diffusion approach on $\mathbb{R}^n$ for general density p to the manifold setting. The majority of this thesis shall be based upon the foundations laid in this subsection. We refer to the preliminaries in Chapter 2 for definitions relating to Riemannian manifolds.

Let M be a Riemannian manifold with metric g. We extend the OU process (1.6) by generalizing the terms in the SDE to the manifold setting: ∇ is now the gradient on M; and the driving process dB_t is now a Brownian motion on M, denoted by dB_t^M. Other components of (1.5), such as the inner product and Laplacian must also be changed to accommodate the potentially non-flat geometry.

In [MG16], Mackey and Gorham employed the principle of diffusion coupling to bound the solution to the Stein equation and its derivatives. The idea is to construct a second diffusion, say Y_t, following the same dynamics as X_t; with the distance between X_t and Y_t going to 0 as time goes on. This is a key principle that was used in the extension to manifold in [LLBF22] which shall be detailed in this subsection. Let μ_ϕ denote the measure on M with Radon-Nikodym derivative

$d\mu_\phi \propto e^{-\phi}d\text{vol}$ where $d\text{vol}$ is the volume form on M. Similarly to the situation in $\mathbb{R}^n$, ϕ is a non-negative function in $C^2(M)$.

Let $\{X_t\}_{t\in\mathbb{R}^+}$ be the M-valued diffusion

$$dX_t = dB_t^M - \frac{1}{2}\nabla\phi(X_t)dt, \quad X_0 = x.$$

The infinitesimal generator of such a diffusion is

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

where Δ_M is the Laplace-Beltrami operator on M. This is indeed a Stein operator, and one can show via integration by parts that $\mathbb{E}_{\mu_\phi}[\mathcal{A}f] = 0$. The Stein equation can then be formulated, and its solution is exactly that of (1.9).

The coupling that was used in [LLBF22] is similar to the Kendall coupling in [Ken86] with the difference being that we do not reflect the direction of the vector field after the parallel transport. The original driving process X_t, dB_t^M, is transported in parallel across a geodesic on M in some direction V_t:

$$dY_t = \Pi_{X_t,V_t}dB_t^M - \frac{1}{2}\nabla\phi(Y_t)dt, \quad Y_0 = y. \tag{1.10}$$

The notation $\Pi_{x,v}u$ indicates that the vector field u has been transported in parallel from a point x in the direction v.

The dynamics of this new diffusion are identical to that of X_t — the infinitesimal generators are the same as well as the invariant distributions. This is because one may rewrite the SDE for Y_t as

$$dY_t = dB_t^{\prime M} - \frac{1}{2}\nabla\phi(Y_t)dt$$

where $dB_t^{\prime M}$ is a Brownian motion on the manifold. The identicalness of the generators follows immediately.

A condition reminiscent of the κ-strongly concave condition in $\mathbb{R}^n$ is needed to ensure that the distance between X_t and Y_t goes to 0. This condition is reformulated on the manifold by taking into account the geometry of M and is summarised in the following theorem [LLBF22].

Theorem 1.1.6. *Let $\phi \in C^2(M)$ be non-negative. Moreover, assume that ϕ satisfies*

$$\mathrm{Ric} + \mathrm{Hess}^{\phi} \geq 2\kappa g. \tag{1.11}$$

Then, the Riemannian distance ρ between X_t and Y_t satisfies

$$\rho(X_t, Y_t) \leq \rho(x, y)e^{-\kappa t}.$$

The notation $\mathrm{Ric} + \mathrm{Hess}^{\phi} \geq 2\kappa g$ means that for any vector field X on TM, $\mathrm{Ric}(X, X) + \mathrm{Hess}^{\phi}(X, X) \geq 2\kappa g(X, X)$. The condition ensures that the tensor on the left hand side of this inequality is positive definite.

Remark. One can recover the original log-concave condition in [MG16] by simply setting $M = \mathbb{R}^n$. $\mathbb{R}^n$ is a flat manifold, and so the sufficient condition becomes $\mathrm{Hess}^{\phi}(v, v) \geq 2\kappa\langle v, v\rangle$. More recently, Gorham et. al. [GDVM19], have refined their argument to allow for the inclusion heavier tailed distributions (distantly dissipative distributions). This is achieved by bounding the Wasserstein metric above by the diffusion Stein discrepancy. As it stands, the Bakry-Èmery-Ricci criterion (1.11) is required (even on compact manifolds) for this method to work.

This powerful result aids one in bounding the solution to the Stein equation and its derivatives, such as:

$$\|f_h\|_\infty \leq \frac{1}{\kappa}\sup_{x\in M}\mathbb{E}[\rho(x, X)], \quad \|df_h\|_\infty \leq \frac{C_0(h)}{\kappa},$$
$$\left\|\mathrm{Hess}^{f_h}\right\|_{\mathrm{op}} \leq \frac{C_1(h)}{2\kappa} + \frac{C_0(h)}{2\kappa^2}C_0((\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#}).$$

By using these bounds on the solution, we can obtain an upper bound on the Wasserstein metric:

Theorem 1.1.7. *Let X and Y be two distributions on M with respective densities μ_ϕ and μ_ψ. Assume that both densities satisfy the condition* (1.11)*. Then the Wasserstein distance between X and Y is bounded above:*

$$d_W(X,Y) \leq \frac{1}{2\kappa}\mathbb{E}[|\nabla(\phi-\psi)|(X)].$$

Remark. A different formulation for Stein's method on manifolds was independently developed by Thompson in [Tho20] which instead uses the theory of martingales to obtain properties of the solution to the Stein equation. The coupling detailed above is absent and instead Thompson calculates bounds on the derivatives of the solution to the Stein equation directly, using damped stochastic parallel translation — see Section 5.5. One of the main difference in results is with the second derivative. Their bound on the second derivative is

$$\left\|\mathrm{Hess}^{f_h}\right\|_{\mathrm{op}} \leq c_1 C_0(h).$$

Their bound is only dependent on the first derivative of h, $C_0(h)$ whereas the bound in [LLBF22] depends on both $C_0(h)$ and $C_1(h)$. It is unclear, however, how c_1 can be obtained in terms of the assumptions of the Bakry-Èmery-Ricci criterion (1.11).

Remark. Literature on the formulation, solution and bounding of the solution of the Poisson equation on manifold

$$\frac{1}{2}\Delta_M u_h - \frac{1}{2}g(\nabla\phi, \nabla u_h) = \bar{h}$$

using PDE theory is quite sparse; [Aub12] provides existence of the solution and bounds to Greens functions inside the solution for compact manifolds [Aub12, Theorem 4.7], and [MSW19] presents existence and Greens function estimates for complete manifolds, however their restrictions on ϕ are more demanding than those in [LLBF22], see [MSW19, Theorem 1.9] which, for example, require the Bakry-

Èmery-Ricci tensor to be bounded from below, oscillation of ϕ to be bounded in a unit ball and the spectral gap of the operator $\mathcal{A}$ to be positive.

1.2 Structure of the Thesis

The remainder of the thesis is structured as follows:

Chapter 2 shall outline important concepts and definitions that are required for this work. We briefly overview the necessary aspects of Riemannian geometry, stochastic differential equations and Brownian motion on manifolds.

When looking at the work of [LLBF22], the framework of Chapter 4, it became clear that many distributions on $\mathbb{S}^1$ could not be used in the context of the diffusion approach. Chapter 3 concerns the extension of Stein's method to the unit circle $\mathbb{S}^1$, with emphasis on the von-Mises distribution. The method used is an adaptation of the previously developed density method, where the Stein kernel and operators have been redefined to reflect the geometry of $\mathbb{S}^1$. An upper bound on the Wasserstein metric is presented in Theorem 3.3.2 and examples follow.

Chapter 4 builds upon the work in [LLBF22] by applying it to several examples. More concretely, it is devoted to constructing explicit analytic bounds on the Wasserstein metric between two probability measures on a number of Riemannian manifolds using the approach detailed in Section 1.1.3. We shall use Theorem 4.0.1 to achieve such. We explore the spaces $\mathbb{S}^n$, $\mathbb{H}^3$, $\mathrm{SO}(n)$, and $\mathcal{P}_n$ (the space of $n \times n$ symmetric positive definite matrices) and try to provide interpretations of results and their consequences when possible; this is particularly the case on $\mathbb{H}^3$ where a finite parameter proof of the Varadhan asymptotic relation is presented.

Chapter 5 is the main technical part of the work where we extend the framework for Stein's method on Riemannian manifolds developed in [LLBF22] and built upon in Chapter 4 to the case where the manifold has a boundary. The general idea is the same as the base manifold case, however a local time term is inserted into the Stochastic Differential Equation to act as a reflecting component.

The addition of this reflecting component presents new challenges for extending the framework, particularly in defining the notion of an invariant measure and in bounding the Lipschitz constant of the second derivative of the solution to the Stein equation. We find that some results from the base manifold case are the same when the local time term is also present, and that results obtained in the manifold without boundary case can be retained by setting $\partial M = \emptyset$.

We conclude the work in Chapter 6 by reiterating our main contributions and findings, along with a discussion of future directions of research.

There are 4 appendices in total, A through D, which give some extra background that isn't particularly needed to understand the general theory we discuss. They do, however, give more detail on certain concepts that are not discussed enough to warrant their own section in the preliminaries.

Chapter 2

Preliminaries

We devote this chapter to reviewing known results and definitions that will be required throughout this work. We first begin with a brief introduction to Riemannian geometry, curvature and the exponential map. We then move over to the stochastic side by defining stochastic differential equations in general, the semigroup and end the chapter with a construction of Brownian motion on manifolds.

2.1 Riemannian Geometry

This section is primarily made up of definitions and results from the following references: Cheeger & Ebin [CE08]; Jöst [Jos08]; Gallot, Hulin, Lafontaine [GHL90]; and Hsu [Hsu02].

Definition 2.1.1. An n-dimensional *topological manifold* M is a Hausdorff, second countable, topological space such that every open neighbourhood in M is homeomorphic to a subset of $\mathbb{R}^n$.

The final statement has some physical intuition; we say that locally, a topological manifold looks like $\mathbb{R}^n$. That is, when zoomed in close enough, the geometry of M begins to mimic the geometry of $\mathbb{R}^n$.

By definition, a manifold with this description has no boundary. Manifolds with a boundary shall be the centrepiece of Chapter 5.

Definition 2.1.2. An n-dimensional topological manifold M is called a *differentiable (or smooth) manifold* if given an atlas $\{U_\alpha, \phi_\alpha\}_{\alpha \in I}$ on M, the transition map $\phi_\beta \circ \phi_\alpha^{-1} : \phi_\alpha(U_\alpha \cap U_\beta) \to \phi_\beta(U_\alpha \cap U_\beta)$ is C^∞ with C^∞ inverse, for all $\alpha, \beta \in I$.

For instance; $\mathbb{S}^n$, the n-sphere requires a minimum of 2 coordinate charts to fully describe the manifold. The n-dimensional Hyperbolic space $\mathbb{H}^n$ only requires 1 chart. In this case we say that there is a global chart that describes the manifold.

A C^k-differentiable manifold can be defined similarly where the transition maps are instead C^k differentiable and not C^∞ differentiable. All manifolds that are discussed in this work will be smooth manifolds. Charts allow us to define local coordinates ϕ_α^{-1} in a coordinate chart (open set) U_α. Local coordinates are frequently used in the examples in Chapter 4.

Definition 2.1.3. The *tangent space* to a differentiable manifold M at a point $x \in M$ is denoted by T_xM and is defined as the vector space over $\mathbb{R}^n$ that is spanned by the partial derivative operators $\frac{\partial}{\partial x_1}, ..., \frac{\partial}{\partial x_n}$.

We may extend this definition to provide a global tangent space, known as the tangent bundle, which contains information about the point and its tangent space.

Definition 2.1.4. The *tangent bundle* (M, TM, π) is expressed as the disjoint union

$$TM := \bigsqcup_{x \in M} T_xM,$$

of all tangent spaces of M. The canonical projection $\pi : TM \to M$ is $\pi(x, v) = x$ for $(x, v) \in TM$.

We denote by $\Gamma(TM)$ the smooth sections of the tangent bundle, by this we mean the smooth map $s_x : M \to T_xM$ such that $\pi \circ s = \text{Id}$.

Definition 2.1.5. Let $f, g \in C^\infty(M)$. Let $X : C^\infty(M) \to C^\infty(M)$ be a linear map that satisfies the Leibniz property $X(fg) = fX(g) + gX(f)$. Then v is called a vector field.

There are three equivalent definitions of vector fields however we take this one for simplicity.

Since the tangent space is a vector space, it is clear that calculus with the partial derivative operators $\{\partial_i\}$ is much easier to perform that doing calculus directly on the manifold. Such a task will require the use of charts, and keeping track of derivatives of overlapping coordinate charts is an arduous task.

Definition 2.1.6. A *Riemannian metric* $g : TM \times TM \to \mathbb{R}$ on M is a symmetric, strictly positive bilinear form on M.

Definition 2.1.7. A smooth manifold equipped with a Riemannian metric is called a *Riemannian manifold.*

Riemannian manifolds will be the spaces of interest throughout this thesis. We shall always equip a manifold with a metric g. By definition, a Riemannian manifold is the couple (M, g), however since M being Riemannian is implied here, we shall simply write that M is a Riemannian manifold and define the metric g when necessary.

Definition 2.1.8. The *Levi-Civita connection* D, is a unique connection on M which is compatible with the metric,

$$D_Z g(X, Y) = g(D_Z X, Y) + g(X, D_Z Y),$$

and is torsion free,

$$T(X, Y) := D_X Y - D_Y X - [X, Y] = 0,$$

for $X, Y, Z \in \Gamma(TM)$.

Definition 2.1.9. The *Riemannian distance* $\rho(x, y)$ between $x, y \in M$ is defined as

$$\rho(x, y) := \inf_{\gamma} \int_0^1 |\dot{\gamma}(t)| dt,$$

where the infimum is taken over all piecewise smooth curves $\gamma : [0,1] \to M$ with $\gamma(0) = x$ and $\gamma(1) = y$.

The notation $\dot{\gamma}$ above is the vector field generated by the curve γ, $\dot{\gamma}(t) = \frac{d\gamma(t)}{dt}$

Definition 2.1.10. The curve or curves which minimize the distance between x and y are named *geodesics* and they satisfy the equation $D_{\dot{\gamma}}\dot{\gamma} = 0$.

Definition 2.1.11. The *parallel transport* $\Pi_{\gamma_{x,y}} : T_xM \to T_yM$ translates a vector $X \in T_xM$ to T_yM along the geodesic γ connecting x and y in such a way that

$$g(X, \dot{\gamma}(0)) = g(\Pi_{\gamma_{x,y}}X, \dot{\gamma}(1)).$$

In other words, the angle between X and $\dot{\gamma}$ is conserved along the whole of γ. Any such vector field X satisfies the equation $D_{\dot{\gamma}}X = 0$.

Parallel transport is a significant tool that is used in the construction of the coupling for the Stein's method in [LLBF22], see Equation (1.10). Such use is also found in Section 5.3 for the same task.

Definition 2.1.12. The *cotangent space* $T_x^*M = (T_xM)^*$ at a point $x \in M$ is the dual space of the tangent space T_xM. It is the space of linear functions on T_xM and elements take the form $\omega = \omega_i dx^i$.

The *cotangent bundle* is defined in a similar way to the tangent bundle. It is the disjoint union

$$T^*M := \bigsqcup_{x \in M} T_x^*M$$

with base space M and canonical projection π.

A section $\omega \in \Gamma(T^*M)$ of the cotangent bundle is called a 1-form on M. The contraction of a 1-form by a vector field yields a scalar; symbolically, $\theta : TM \to \mathbb{R}$. For example, if $\omega = dx^1$ and $X = \partial_1$, then the contraction $\omega(X) = dx^1\partial_1 = 1$. A more general result is that the contraction $dx^i\partial_j = \delta^i_j$, the Kronecker delta tensor.

Remark. Here we have used the Einstein index notation to write out the form by dropping the sum. The contraction of the index i in $\omega_i dx^i$ means that summation over index i is implied.

One can combine notions of vector fields and 1-forms to create a more general object called a tensor.

Definition 2.1.13. The bundle of (r, s)-tensors is the disjoint union

$$T^{r,s}M = \bigsqcup_{x \in M} (T_xM)^{\otimes r} \otimes_{\mathbb{R}} (T_x^*M)^{\otimes s}.$$

An element of the section $\Gamma(T^{r,s}M)$ is called an (r, s)-tensor and some examples include the Riemannian metric g is a (0,2)-tensor and the Kronecker delta is a (1,1)-tensor. If $\{X_i\}$ is an orthonormal frame of TM, then an orthonormal frame in T^*M is denoted $\{dx^i\}$. Elements of a section of $T^{r,s}M$ are uniquely written as

$$\theta = \theta^{i_1 \dots i_r}_{j_1 \dots j_s} X_{i_1} \otimes \dots \otimes X_{i_r} \otimes dx^{i_1} \otimes \dots \otimes dx^{i_s}.$$

In the special case of $r = 0$, the space of p-forms generates an algebra $\Lambda^p(\mathbb{R}^n)$ along with the antisymmetric wedge product $\wedge$.

Definition 2.1.14. The *exterior derivative* d is an operator which increases the degree of a differential form by 1, $d : \Lambda^p \to \Lambda^{p+1}$ with the special property that $d^2 = 0$.

Definition 2.1.15. The *interior product* ι is an operator which decreases the degree of a differential form by 1, $\iota_X : \Lambda^p \to \Lambda^{p-1}$ in which $X \in \Gamma(TM)$.

The interior product is essentially the contraction $\omega(X)$. Similarly to the exterior derivative, it is the case that $\iota_V \circ \iota_W = 0$, $V, W \in \Gamma(TM)$.

Remark. When writing out forms, we shall be dropping the tensor product or wedge product and assume that it is there without explicitly writing it.

The exterior derivative and 1-forms make an crucial appearance in Section 5.5 when we bound the second derivative of the solution to the Stein equation.

The *gradient operator* on a function ∇f is the dual of the 1-form df. It is the unique vector field defined by the relation

$$g(\nabla f, X) = df(X), \quad X \in \Gamma(TM).$$

The *divergence operator* $\nabla \cdot$ is the contraction of the (1,1)-tensor ∇X. By defining the gradient and divergence operators on M, the manifold equivalent to the Laplace operator can also be found. The operator $\Delta_M = \nabla \cdot \nabla$ is known as the *Laplace-Beltrami operator.*

The *Hessian* is the (0,2)-tensor of second order derivatives. For a function $f \in C^\infty(M)$,

$$\text{Hess}^f := D^2 f = Ddf.$$

It is also true that $\Delta_M f = \text{TrHess}^f = \sum_{i=1}^n D^2 f(X_i, X_i)$ for an orthonormal basis $\{X_i\}$ on T_xM. The Hessian can be alternatively defined using curves as

$$\text{Hess}^f(X, X) = \frac{d^2}{dt^2} f(\gamma(t))\bigg|_{t=0}, \tag{2.1}$$

for which $\gamma(0) = x$ and $\dot{\gamma}(0) = X \in T_xM$.

Due to its appearance in the Bakry-Èmery-Ricci criterion 1.11, we shall be frequently calculating the Hessian of many different functions on a number of different spaces in Chapter 4.

The *volume form* is an n-form (matching the dimension of the manifold) defined as

$$d\text{vol} := \sqrt{\det(g)} dx_1 \wedge dx_2 \wedge ... \wedge dx_n.$$

We have written $\det(g)$ as the determinant of the metric when viewed in the matrix form g_{ij}. The volume form allows us to integrate on M. The form $dx_1 \wedge ... \wedge dx_n$ can be regarded as the Lebesgue measure on $\mathbb{R}^n$, meaning that the volume form is a measure, specifically a normalized Hausdorff measure. The volume form is thought of as the uniform measure on M. Further probability measures are constructed by using a Radon-Nikodym derivative. For example, one repeatedly used measure in this work is a measure μ_ϕ which has Radon-Nikodym derivative

$$d\mu_\phi = \frac{1}{C_\phi} e^{-\phi} d\text{vol}.$$

The constant C_ϕ ensures μ_ϕ is a probability measure.

With integration clearly defined on our manifolds, L^2 spaces can be constructed and operators and their duals can be defined in a more concrete sense. The Laplace-Beltrami operator, like its Euclidean counterpart, is a self-adjoint operator on $L^2(M, d\text{vol})$. It is a particularly important operator in the construction of Brownian motion on M.

The integration by parts formula (or Green's formula) on M with $\partial M = \emptyset$ is

$$\int_M g(\nabla f, X) d\text{vol} = -\int_M f \nabla \cdot X d\text{vol}.$$

When $X = \nabla h$ for some function $h \in C^\infty(M)$, this becomes

$$\int_M f \Delta_M h + g(\nabla f, \nabla h) d\text{vol} = 0.$$

2.1.1 Curvature

Curvature is a defining quality which distinguishes manifolds from each other. As is usual, let M be an n-dimensional, complete, connected Riemannian manifold with metric g and Levi-Civita connection D.

Definition 2.1.16. Let $X, Y, Z \in \Gamma(TM)$, the *Riemannian curvature tensor* is defined as

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

The final term involving the Lie bracket is identically 0 if X and Y are independent vector fields and so the Riemannian curvature is sometimes written in the form

$$R(X,Y)Z = [D_X, D_Y]Z.$$

The *sectional curvature* is related to the Riemannian curvature by

$$K(X,Y) := g(R(X,Y)Y, X).$$

Geometric analysis on manifolds is typically categorised into three sections, manifolds with positive, non-positive, and non-negative sectional curvature which is why sectional curvature is a key characteristic of a manifold.

Definition 2.1.17. The *Ricci curvature* tensor, or just Ricci curvature for short, is a contraction of the Riemannian curvature, defined as

$$\mathrm{Ric}(X,Y) := \sum_{i=1}^{n} g(R(X,X_i)X_i, Y),$$

where $\{X_i\}_{i=1}^n$ is an orthonormal basis $\pi(X) = \pi(Y)$.

We may alternatively rewrite this in terms of the sectional curvature,

$$\mathrm{Ric}(X,X) = \sum_{i=1}^{n} K(X,X_i).$$

The last curvature quantity we define is the *scalar curvature*, a contraction of the Ricci tensor,

$$S := \sum_{i=1}^{n} \mathrm{Ric}(X_i, X_i),$$

which is a scalar quantity and cannot be further contracted with the metric.

The Ricci curvature is a contributor to the Bakry-Èmery-Ricci criterion (1.11) and is pivotal in determining whether the diffusion approach in [LLBF22] can be used.

Einstein manifolds are manifolds whose Ricci curvature is proportional to the metric. Explicitly, $\mathrm{Ric} = \lambda g$ for some $\lambda \in \mathbb{R}$. If a manifold has constant sectional curvature, i.e. $K(X,Y)=K$ then it is an Einstein manifold with $\lambda = K(n-1)$. For example,

- For $M=\mathbb{S}^n$, $K=1$ and $\mathrm{Ric}=(n-1)g$,
- For $M=\mathbb{R}^n$, $K=0$ and obviously $\mathrm{Ric}=0$,
- for $M=\mathbb{H}^n$, $K=-1$ and $\mathrm{Ric}=-(n-1)g$.

In these special cases, it is unnecessary to calculate the Ricci curvature in local coordinates and we frequently use these facts in Section 4.1 for $\mathbb{S}^n$ and Section 4.2 for $\mathbb{H}^n$ when verifying the Bakry-Èmery-Ricci criterion.

2.1.2 Exponential Map

In Definition 2.1.10 it was described that geodesics may not always be unique. For example, when taking $M=\mathbb{S}^n$, and $x=N$, $y=S$, the north and south poles respectively, there is more than one geodesic connecting x and y — in fact there are infinitely many. We define the cut locus of a point $x \in M$ as

$$\mathrm{cut}_x = \{y \in M : \text{there is no unique minimizing geodesic connecting } x \text{ and } y\}.$$

For an example of such a set, take $M=S^1$, if we set x to be the north pole, then cut_x is the south pole. A subset of the cut locus, the conjugate points, are the set of points such that there are infinitely many geodesics emanating from x to y. For example, any antipodal point of a point in $\mathbb{S}^n$, $n>1$ is a conjugate point.

Questions about injectivity of geodesics give rise to the exponential map.

Definition 2.1.18. Let M be a Riemannian manifold, $x \in M$, $v \in T_xM$, and $\gamma : [0,1] \to M$ a geodesic with $\gamma(0)=x$. Define

$$V_x := \{v \in T_xM : \dot{\gamma}(0)=v \text{ and } \gamma(1) \notin \mathrm{cut}_x\}.$$

Then the mapping $\exp_x : V_x \to M$, $v \mapsto \gamma(1)$ is called the *exponential map* of M at x.

Explicitly, for a geodesic with $\gamma(0) = x$ and $\dot{\gamma}(0) = v$, $\exp_x(v) = \gamma(1)$ assuming that v is unique. The injectivity radius is then defined as

$$\operatorname{inj}(M) = \min_x \sup\{|v| : x \in M, v \in T_xM,\ \exp_x(v) \text{ is injective}\}.$$

Note that the minimum is taken over the whole manifold so that regardless of the point we take, we never enter the cut locus.

Definition 2.1.19. We say $y = \gamma(1)$ is conjugate to $x = \gamma(0)$ along γ if $d\exp_x|_{\dot{\gamma}(0)}$ is not full rank.

We can similarly define the inverse of the exponential map which instead maps a neighbourhood of p to the tangent space at x, $\exp_p^{-1} : U \to T_xM$. This definition of exponential inverse allows us to show that the exponential map is a local diffeomorphism. The exponential mapping $\exp_x$ maps a neighbourhood of $0 \in T_xM$ diffeomorphically onto a neighbourhood of $x \in M$. For manifolds with non-positive sectional curvature, this mapping is diffeomorphic globally on the manifold by the well known Cartan-Hadamard theorem. Moreover, the theorem also tells us that the exponential map is a covering map for the manifold. Locally, one may think of the tangent space around 0 as Euclidean space, and so the ordered pair $(U, \exp_x^{-1})$ is infact a chart to $\mathbb{R}^n$. It is also true that for manifolds of non-positive sectional curvature, no point has a corresponding conjugate point.

2.2 Stochastic Differential Equations on $\mathbb{R}^n$

Owing to the name ‘diffusion approach’ for Stein’s method, stochastic differential equations (SDEs) are central to the framework of the approach. The underlying mathematics that allows Chapter 4 to work relies on the use of SDEs. In Chapter 5, SDEs are also used, albeit with a twist in the form of a reflection term.

For this section, we have consistently referred to: Rogers and Williams volumes 1 & 2, [RW94, RW00]; Ikeda & Watanabe [IW14]; Øksendal [Øks13]; and Hsu [Hsu02].

In this section, $\{X_t\}_{t\in\mathbb{R}^+}$ is a stochastic process on filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in\mathbb{R}^+}, P)$ which takes values in $\mathbb{R}^n$ (and later a Riemannian manifold).

Definition 2.2.1. A process $\{X_t\}_{t\in\mathbb{R}^+}$ is called a *Markov process* if it satisfies the Markov property

$$\mathbb{E}_x[f(X_{t+s})|X_s] = \mathbb{E}_{X_s}[f(X_t)], \quad \text{a.s.}, \quad X_0 = x$$

for any bounded, measurable function $f : \mathbb{R}^n \to \mathbb{R}$.

We denote by P_t the transition functional acting on bounded measurable functions

$$P_t f(x) = (P_t f)(x) := \int_{\mathbb{R}^n} f(y) P_t(x, dy),$$

where $P_t(x, y)$ is the transition kernel. This gives rise to the *Chapman-Kolmogorov equation*

$$P_s P_t = P_{s+t}, \quad s, t \geq 0.$$

The transition functional P_t is more commonly known as the *semigroup* due to the Chapman–Kolmogorov equation. The differential $P_t(x, dy)$ is named the *transition density function* and is written $P_t(x, dy) = p(t, x, y)dy$, assuming that the Lebesgue measure and transition density are absolutely continuous — we can write it as a Radon-Nikodym derivative. The heat kernel is the name of the distribution that has probability density function equal to that of the transition density function of a Brownian motion for a fixed initial point. The heat kernel of $\mathbb{H}^3$ is one of the two distributions that we compare in Section 4.2.

Definition 2.2.2. A *Feller-Dynkin* (FD), or sometimes just *Feller*, semigroup is a strongly continuous, contraction, Markov semigroup $\{P_t\}_{t\in\mathbb{R}^+}$ of linear operators on $C_0(\mathbb{R}^n)$ (the space of continuous functions on $\mathbb{R}^n$ that vanish at infinity).

In other words,

i) $P_t : C_0(\mathbb{R}^n) \to C_0(\mathbb{R}^n)$,

ii) $\forall f \in C_0(\mathbb{R}^n)$, $\|P_t f - f\|_\infty \to 0$ as $t \to 0^+$

iii) $\forall f \in C_0(\mathbb{R}^n)$, $0 \leq f \leq 1 \implies 0 \leq P_t f \leq 1$,

iv) $P_s P_t = P_{s+t}\ \forall s, t \geq 0$, $P_0 = \mathrm{Id}$,

Definition 2.2.3. A *Feller process* is a Markov process with Feller semigroup.

The semigroup property and strong continuity of the Feller semigroup suggests that we should have (in some manner)

$$\frac{d}{dt} P_t = \lim_{s \to 0} \frac{P_{t+s} - P_t}{s} = P_t \lim_{s \to 0} \frac{P_s - \mathrm{Id}}{s} = P_t \mathcal{A} = \mathcal{A} P_t$$

in which Id is the identity operator $\mathrm{Id} \circ f = f$ and

$$\mathcal{A} := \lim_{s \to 0} \frac{P_s - \mathrm{Id}}{s}.$$

The quantity $\mathcal{A}$ is the *infinitesimal generator* of the Markov process X_t.

Definition 2.2.4. An *Itô diffusion* on $\mathbb{R}^n$ is a process $\{X_t\}_{t \in \mathbb{R}^+}$ that satisfies the SDE

$$dX_t = \sigma(X_t) dB_t + b(X_t) dt,$$

where $\{B_t\}_{t \in \mathbb{R}^+}$ is an $\mathbb{R}^n$-valued Brownian motion, $b : \mathbb{R}^n \to \mathbb{R}^n$ and $\sigma : \mathbb{R}^n \to \mathbb{R}^{n \times n}$.

The SDE of the Itô diffusion X_t has a unique, strong solution if we have globally Lipschitz coefficients;

$$|b(x) - b(y)| + \|\sigma(x) - \sigma(y)\| \leq C|x - y|$$

for some constant $C > 0$ for all $x, y \in \mathbb{R}^n$. Itô diffusions have many useful properties. They satisfy the Markov property, and even further, have a Feller semigroup. Because of this, the infinitesimal generator exists and can be used to characterise each individual Itô diffusion. For example, Let $X_t = B_t$, an $\mathbb{R}^n$ Brownian motion. Then the infinitesimal generator $\mathcal{A}f(x) = \frac{1}{2}\Delta f$. For a general SDE with drift b and diffusion matrix σ, the infinitesimal generator is

$$\mathcal{A}f(x) = \langle b(x), \nabla f(x)\rangle + \frac{1}{2}\mathrm{Tr}(\sigma^\intercal(x)\nabla^2 f(x)\sigma(x)).$$

Definition 2.2.5. A Borel measure $\mu(dx)$ on $\mathbb{R}^n$ is called an *invariant measure* of an Itô diffusion with semigroup P_t if

$$\int_{\mathbb{R}^n} P_t f(x)\mu(dx) = \int_{\mathbb{R}^n} f(x)\mu(dx), \quad \forall f \in C_0^2(M).$$

If a random variable $X \sim \mu$, we may rewrite this condition in terms of the infinitesimal generator as $\mathbb{E}_\mu[\mathcal{A}f(X)] = 0$. An invariant measure is called a stationary distribution if $\mathbb{E}_\mu[1] = 1$. The definition of the invariant measure in terms of the infinitesimal generator will be the principal way that we shall find a Stein operator for a given probability measure.

The Itô integral is constructed by taking the left end point of the integrand in the Riemann-Stieltjes integral, and then taking the limit in the L^2 sense;

$$\int_0^t f(s)dB_s = \lim_{\mathrm{mesh}(P)\to 0} \sum_{i=1}^n f_{t_{i-1}}(B_{t_i} - B_{t_{i-1}}), \quad \text{in } L^2.$$

The resulting Itô integral has many desirable properties, particularly the martingale property. However, if we were instead to take the midpoint, we arrive at the *Stratonovich integral.* We differentiate between Itô and Stratonovich integrals by using $\circ$, e.g.

$$\int_0^t f(s) \circ dB_s = \lim_{\mathrm{mesh}(P)\to 0} \sum_{i=1}^n \frac{f_{t_i} + f_{t_{i-1}}}{2}(B_{t_i} - B_{t_{i-1}}), \quad \text{in } L^2.$$

The Stratonovich integral is much more familiar in the sense that it obeys the classical rules of calculus such as the chain rule. We later see in the subsequent section, Section 2.3, that the Stratonovich integral is a more natural framework to construct a Brownian motion on manifold than it is to use the Itô integral.

Lemma 2.2.6 (Itô Lemma)**.** *Let X_t be the Itô diffusion that is the solution to the SDE $dX_t = b(X_t)dt + \sigma(X_t)dB_t$. Then for $f \in C^2(\mathbb{R}^n)$,*

$$df(X_t) = \langle \nabla f(X_t), dX_t \rangle + \frac{1}{2} dX_t^{\intercal} \nabla^2 f(X_t) dX_t.$$

One can convert from Itô to Stratonovich form and vice versa by using the following conversion formula:

$$\int_0^t \sigma(X_s) \circ dB_s = \int_0^t \sigma(X_s) dB_s + \frac{1}{2} \int_0^t \nabla_{\sigma_j} \sigma_i(X_s) d\langle B^i, B^j \rangle_s,$$

where we have used the notation that σ_i is the i-th row of σ and $\nabla_{\sigma_j}\sigma_i$ is the directional derivative of σ_i in the direction of σ_j.

2.3 Brownian Motion on Riemannian manifolds

We now discuss the intrinsic construction of Brownian motion on manifolds. For this section, we have primarily used the excellent exposition in Hsu [Hsu02]. Appendix C has been written to provide an introduction to the orthonormal frame bundle.

To be able to extend Brownian motion to manifolds it makes sense to map our Brownian motion on $\mathbb{R}^n$ onto the manifold. However, maps straight from $\mathbb{R}^n$ to M are difficult to obtain in general and some charts, for example $\mathbb{S}^2$, mappings to $\mathbb{R}^n$ will only be true for a subset of the whole space. Another reason as to why a direct method will not work is as follows: An M-valued Brownian motion should have the general form

$$dX_t = V_\alpha(X_t) \circ dB_t^\alpha$$

for some set of vector fields $V_i, i = 1, ..., n$ on TM. This process has the following infinitesimal generator;

$$\mathcal{A} = \frac{1}{2}\sum_{i=1}^{n} V_i^2.$$

However, there is no way that we can globally write the generator $\mathcal{A}$ as a sum of squared non-vanishing vector fields for general M. For example, on $\mathbb{S}^n$, we have the Hairy Ball Theorem, which tells us precisely this.

An intrinsic formulation for Brownian motion on M is known as the Eells-Elworthy-Malliavin construction and is described in the following procedure. We must first find a space on which we can write the Laplacian as a sum of squared vector fields. The precise bundle with which this is achieved is the horizontal bundle H. Let $\mathcal{O}(M)$ be the orthonormal frame bundle and define Böchner's horizontal Laplacian

$$\Delta_{\mathcal{O}(M)} = \sum_{i=1}^{n} H_i^2, \tag{2.2}$$

where H_i are the fundamental horizontal vector fields (shortened from H_{e_i}). The relation of Bochner's horizontal Laplacian to the Laplace-Beltrami operator is established in the following lemma:

Lemma 2.3.1. *Let $f \in C^\infty(M)$ and define $\tilde{f} = f \circ \pi$ the lift from M to $\mathcal{O}(M)$. Then for any $u \in \mathcal{O}(M)$,*

$$\Delta_M f(x) = \Delta_{\mathcal{O}(M)} \tilde{f}(u),$$

with $x = \pi u$.

This key link gives us the necessary tools to construct a Brownian motion on $\mathcal{O}(M)$ and then map down to M, giving rise to an M-valued Brownian motion.

Consider the following Stratonovich SDE

$$dU_t = H(U_t)_\alpha \circ dB_t^\alpha, \quad U_0 = u = \pi^{-1}(x),$$

where H is the horizontal lift of $(U_t e)^*$ and B_t is an n-dimensional Brownian motion. Then U_t is a Brownian motion on $\mathcal{O}(M)$. Recognise that the infinitesimal generator of this Stratonovich process is $\frac{1}{2}\sum_{\alpha=1}^{n} H_{\alpha}^{2}$, or in other words, $\frac{1}{2}\Delta_{\mathcal{O}(M)}$. Since we confirm that U_t is a Brownian motion, it must obey the Itô formula

$$F(U_t) = F(U_0) + \sum_{\alpha=1}^{n} \int_0^t H_{\alpha} F(U_s) dB_s^{\alpha} + \frac{1}{2}\int_0^t \Delta_{\mathcal{O}(M)} F(U_s) ds,$$

for some smooth function F on $\mathcal{O}(M)$. If we now take $F = f \circ \pi$, for another smooth f on M and defining $X_t = \pi U_t$, we have that

$$f(X_t) = f(X_0) + \sum_{\alpha=1}^{n} \int_0^t H_{\alpha} f(X_s) dB_s^{\alpha} + \frac{1}{2}\int_0^t \Delta_M f(X_s) ds$$

using Lemma 2.3.1. We can now infer that X_t has infinitesimal generator $\frac{1}{2}\Delta_M$, and hence, X_t is a Brownian motion on M.

If one does possess local coordinates on M, then it is possible to find an explicit form for the SDE of the Brownian motion on M. The form of the horizontal lift in local coordinates, as calculated in [Hsu02], is $\Xi(x) = g^{-1/2}(x)$, where $g^{-1/2}$ is denoted as the square root of the inverse of the metric tensor g. In terms of an Itô diffusion this is

$$dX_t^i = \sigma_{ij} dB_t^j + \frac{1}{2} b^i dt$$

where $\sigma_{ij} = g_{ij}^{-1/2}$ and $b^i = g^{jk}\Gamma_{jk}^i$. Then

$$\begin{aligned} \mathcal{A}f &= \frac{1}{2} g^{ij}\partial_{ij} f - \frac{1}{2} g^{jk}\Gamma_{jk}^i \partial_i f, \\ &= \frac{1}{2}\Delta_M f, \end{aligned}$$

showing that X_t is indeed a Brownian motion on M.

Example 2.3.2. Let $M = \mathbb{H}^n$ and $(x_1, x_2, ..., x_n)$ be the standard local (global) coordinates for the Poincaré half plane model. The horizontal lift on $\mathbb{H}^n$ is calculated intrinsically, $\Xi(x) = x_n I_n$ where I_n is the $n \times n$ identity matrix. Then a standard Brownian motion on $\mathbb{H}^2$ is constructed via the following Itô SDE:

$$dX_t^i = X_t^n dB_t^i, i = 1, ..., n,$$

where B_t^i is a Brownian motion on $\mathbb{R}$. One may also check that the infinitesimal generator of this diffusion coincides with the Laplacian on $\mathbb{H}^n$.

An important question is to ask whether it is possible for the Brownian motion to exit (explode) the manifold in finite time. An example of this happening on $\mathbb{R}$ is taking the process $dX_t = \frac{1}{X_t} dB_t$, $X_0 \neq 0$. Then the process hits 0, exploding in finite time and leaves $M = \mathbb{R}$. Techniques such as adding coffin states to manifolds is a workaround to this problem, but then the manifold may no longer connected. This gives rise to the concept of *stochastic completeness*. We say that a manifold M is stochastically complete if for every $x \in M$, $P_x(e = \infty) = 1$ where e is the explosion time (the time to leaving the manifold) and P_x represents the Brownian motion starting at the point x. Checks for stochastic completeness are well developed (see [Hsu02, Chapter 4]) and it has been shown that a control on the lower bound of the Ricci curvature is vital to ensure that the Brownian motion does not escape.

Like with SDEs on $\mathbb{R}^n$, we introduce a drift term with the addition of a finite variation process in t:

$$dX_t = dB_t^M - \frac{1}{2}\nabla\phi(X_t)dt,$$

where $\{B_t^M\}_{t\in\mathbb{R}^+}$ is a Brownian motion on M and $\phi \in C^2(M)$. This process is Feller and has infinitesimal generator

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

Such a process does not leave M in finite time [Bak86] provided that there exists a constant $\kappa > 0$ such that

$$\text{Ric} + \text{Hess}^\phi \geq -\kappa g, \quad \forall x \in M.$$

Chapter 3

Stein's Method for Probability Measures on $\mathbb{S}^1$.

This chapter concerns the formulation of a Stein's method for the $\mathbb{S}^1$, the unit circle.

The main object of study in this chapter is what is known as the Stein kernel — not to be confused with the kernels of the Stein kernel discrepancy. This object was first studied by Stein in [Ste86] and has now been extensively studied in both univariate (for example in [LRS17b]) and multivariate [MRS18] cases on Euclidean space. Its use to bound the Wasserstein metric has been established in [LRS17a] along with an application in Bayesian statistics. This theory has yet to be extended to $\mathbb{S}^1$. The problem yet lies in obtaining analytic bounds on the Wasserstein metric using this approach. The two notable approaches to this problem are the diffusion approach, and the classical approach — by classical, we mean in the same spirit as Stein in [Ste86]. For the sake of completeness, we shall give a brief analysis of both in the context when $M = \mathbb{S}^1$ to select a method.

As described in the Introduction, a recent development within the area of Stein's method [LLBF22] enables us to extend the diffusion method originally presented in [MG16] to general manifolds. We shall begin this chapter by verifying that the diffusion approach is inapplicable in the case for most popular circular distributions. Suppose, for example, it was desirable to construct a Stein's method for the von-Mises distribution. The probability density function (pdf) of the von-Mises distribution, $\mathrm{VM}(\mu, \kappa)$, has the form

$$p_{\mathrm{VM}}(x|\kappa, \mu) = \frac{e^{\kappa \cos(x-\mu)}}{2\pi I_0(\kappa)}, \quad x \in \mathbb{S}^1,\ \kappa > 0,\ \mu \in \mathbb{S}^1 \tag{3.1}$$

where $I_n(x)$ is the modified Bessel function of the first kind defined as $I_n(\kappa) := \frac{1}{\pi}\int_0^\pi e^{\kappa\cos\theta}\cos(n\theta)d\theta$. Here we have identified $\mathbb{S}^1$ with the interval $(-\pi, \pi]$. In the discussion paper [Ken75], Kent proposed the following SDE that has the von-Mises distribution as its stationary distribution,

$$dX_t = -\frac{\kappa}{2}\sin(X_t - \mu)dt + dB_t, \tag{3.2}$$

the solution to which is aptly named the von-Mises process. The infinitesimal generator of (3.2) is

$$\mathcal{A}f(x) = \frac{1}{2}f''(x) - \frac{\kappa}{2}\sin(x-\mu)f'(x). \tag{3.3}$$

Using the generator (3.3) one can show — by application of integration by parts — that the invariant measure of the von-Mises process is indeed the von-Mises distribution.

As discussed in Section 1.1.3, for the diffusion approach to be applicable, we impose the sufficient condition on the density function and geometry

$$\mathrm{Ric} + \mathrm{Hess}^{\phi} \geq 2kg, \tag{3.4}$$

for some $k > 0$ and where the density function is written in the form $p(x) \propto e^{-\phi(x)}$. On $\mathbb{S}^1$, this condition is simplified to the $2k$-strongly log-concave assumption $\mathrm{Hess}^{\phi} \geq 2kg$ due to the flat geometry of the circle.

However, this is clearly not satisfied for the von-Mises distribution on $\mathbb{S}^1$. For $\phi = -\kappa\cos(x-\mu)$, $\mathrm{Hess}^{\phi}(x) = \kappa\cos(x-\mu)$ is not strictly positive on $\mathbb{S}^1$ and so the sufficient condition fails to be satisfied in this case for any x, μ, or κ. In summary, we require ϕ to be a convex function, which for many classical circular distributions such as the Bingham uniform, cardioid or wrapped distributions is not true. This motivates the need to use classical methods in order to construct a Stein method for distributions on $\mathbb{S}^1$.

It now seems reasonable to apply the density method discussed in Section 1.1.1 on an interval $[a, b]$ and identify this as the circle. However, it must be appreciated that in order to equate the circle with an interval it is neccessary to assign a wrapping at the endpoints of the interval. This means one can not simply employ general density methods discussed in, for example, [CGS10]. Boundary conditions on f_h and p must be obtained for these methods to be applicable — particularly, $\lim_{x\to a+} f(x)p(x) = \lim_{x\to b-} f(x)p(x)$. The von-Mises, Bingham and uniform distributions do not satisfy this boundary condition unless one restricts the function space for which the operator is defined on. Instead, we will modify the density approach to accommodate the geometry of $\mathbb{S}^1$, and we shall see that by the definition of continuous functions on $\mathbb{S}^1$, this condition is always satisfied for absolutely continuous f and p. Instead of identifying the circle as a wrapping of an interval of arbitrary length, for the purpose of this chapter, we shall identify it with an interval of length 2π.

The Stein kernel, which shall be defined later, has been shown [LRS17a, GL18] to provide an alternative way to construct analytic bounds on the Wasserstein metric between two known distributions. For this reason, we shall be utilising the Stein kernel to bound the Wasserstein metric. This avoids the need to bound the solution to the Stein equation, which can sometimes yield loose bounds. For example, when looking at the von-Mises distribution $X \sim \mathrm{VM}(0, \kappa)$, one will find that the solution to the Stein equation has the form

$$f_h(x) = e^{-\kappa\cos(x)} \int_{-\pi}^{x} (h(u) - \mathbb{E}[h(X)])e^{\kappa\cos(u)} du.$$

This cannot be bounded via conventional means, i.e. using properties of the CDF, due to the oscillatory nature of the cosine function in the exponent. Bounding it directly using the Lipschitz continuity of h will result in very large upper bounds. For example, one way to bound it is

$$\begin{aligned} f_h(x) &= e^{-\kappa\cos(x)} \int_{-\pi}^{x} (h(x) - \mathbb{E}[h(X)]) e^{\kappa\cos(x)} dx \\ &\leq \left\|h'\right\|_\infty e^{-\kappa\cos(x)} \int_{-\pi}^{x} \mathbb{E}[\rho(u,X)] e^{\kappa\cos(u)} du \end{aligned}$$

which can be bounded above by $2e^{2\kappa}\pi^2 \left\|h'\right\|_\infty$. This is a rather loose bound that is not very good in practice.

In addition to this, one usually relies upon applying one or more of: exchangeable pairs, size-biasing, sum of variables and zero-biasing in order to bound the Wasserstein metric. This will not be needed when working with the kernel as our method, reminiscent of [LRS17a], will directly compare the operators of our two distributions to obtain an upper bound.

The chapter is laid out in the following way: Section 3.1 sets the foundation for the Stein operator and its inverse, translating their definitions and properties from $\mathbb{R}$ to $\mathbb{S}^1$. The Stein kernel as described by Stein in [Ste86] will also be introduced. In addition, we shall discuss why this kernel is not suitable for common and widely used distributions on $\mathbb{S}^1$ when performing analysis. This motivates the need to construct a new kernel. In Section 3.2, we shall construct a new kernel, the circular Stein kernel, that one can explicitly calculate for many common circular distributions, e.g. von-Mises and Bingham. This kernel will also satisfy properties akin to those of the classical Stein kernel. With the circular Stein kernel, Section 3.3 discusses a way in which one can use it to bound the Wasserstein distance. Examples of this bound shall also be presented, in particular, we shall exhibit an application in Bayesian statistics and a numerical approximation for the bound between the von-Mises and wrapped normal distributions.

Notation and conventions. Throughout the chapter we shall be using the following notation: P is a probability measure on $\mathbb{S}^1$ with continuous Lebesgue density p. $L^1(P)$ denotes the space of absolutely integrable functions on $\mathbb{S}^1$ under P. We use this abbreviation for $L^1(\mathbb{S}^1, P)$ unless explicitly stated otherwise. For simplicity, we shall assume that the support of P is a connected subset of $\mathbb{S}^1$. Any reference to standard coordinates of $\mathbb{S}^1$ means that we associate $\mathbb{S}^1$ with the interval $(-\pi, \pi]$ alongside the equivalence relation $x \sim y$; meaning that if $x \sim y$, then $x - y \bmod 2\pi = 0$. We prescribe $\mathbb{S}^1$ with its canonical Riemannian metric $g = dx^2$.

3.1 The Circular Stein Operator

This initial section is dedicated to establishing the framework necessary to formulate the Stein equation on $\mathbb{S}^1$. We begin by defining the canonical Stein operator for $\mathbb{S}^1$ and further define its inverse operator. In lieu of a diffusion approach, we shall pursue a modified density approach which draws inspiration from Döbler [Döb15].

We first recall two definitions from analysis:

Definition 3.1.1. A function f is absolutely continuous on $(a, b]$ if f has derivative f' almost everywhere, $f' \in L^1((a, b], dx)$ and one can write

$$f(x) = f(a) + \int_a^x f'(y)dy, \quad a < x \le b.$$

Definition 3.1.2. Let $g \in C^\infty((a, b])$ with $g(a) = g(b) = 0$ be a given function. We say a function $\phi \in L^1((a, b], dx)$ has weak derivative ϕ' if

$$\int_a^b \phi(x)g'(x)dx = -\int_a^b \phi'(x)g(x)dx.$$

3.1.1 The Density Operator

To begin, we start with constructing the foundations of Stein's method for general univariate distributions on $\mathbb{S}^1$:

Definition 3.1.3. Let P be a probability measure on $\mathbb{S}^1$ with Lebesgue density p with the assumption that $p' \in L^1(dx)$. Define $\mathcal{I} = \{x \in \mathbb{S}^1 : p(x) > 0\}$. The Stein class $\mathcal{F}(P)$ of P is the collection of functions $f : \mathbb{S}^1 \to \mathbb{R}$ such that

i) f is differentiable everywhere on $\mathcal{I}$,

ii) $f' \in L^1(dx)$,

iii) $\int_{\mathcal{I}} (fp)' dx = 0$.

Since we assume p is absolutely continuous, it is immediate that for any $f \in \mathcal{F}(P)$ the product fp is absolutely continuous on $\mathcal{I}$ since f is also absolutely continuous by items i) and ii). Because of the Lebesgue integrability assumption on p', constant functions are always in $\mathcal{F}(P)$, and hence $\mathcal{F}(P)$ is always non-empty.

Definition 3.1.4. The Stein operator $\mathcal{T}_p$ of a probability measure P on $\mathbb{S}^1$ is the mapping

$$\mathcal{T}_p : \mathcal{F}(P) \to L^1(P)$$

given by

$$\mathcal{T}_p f(x) = \begin{cases} \frac{(fp)'}{p}(x) & x \in \mathcal{I}, \\ f(x) & x \notin \mathcal{I}. \end{cases} \tag{3.5}$$

Note that we will not be characterising P with the Stein class and Stein operator, this requires one to show that $\mathbb{E}_Q[\mathcal{T}_P f(X)] = 0 \implies Q = P$, we only need the fact that expectation under P is 0 for the framework to hold.

By comparing these definitions with their Euclidean counterparts (for example in [LRS17b]), we see immediate differences. Instead of requiring fp to be absolutely continuous, we may instead restrict $f' \in L^1(dx)$ so that fp is absolutely continuous. In other words, we need not demand absolute continuity in fp since it is a consequence of Definition 3.1.4. Stating that f' is differentiable everywhere allows us to write down, explicitly, the Stein operator as a differential operator in terms of f — see below. The key difference is that for the definition of the Stein class on $\mathbb{R}$, the third condition $\int_{\mathbb{R}} (fp)' dx = 0$ is required (cf. [LRS17b]) for

$p \in L^1(\mathbb{R}, dx)$. In the case of the circle, if $\text{supp}(P) = \mathbb{S}^1$, this condition is automatically satisfied: if we are to identify $\mathbb{S}^1$ with the interval $[-\pi, \pi)$ alongside the equivalence relation defined earlier, then $f(-\pi)p(-\pi) = f(\pi)p(\pi)$ and so

$$\int_{\mathbb{S}^1} (fp)' dx = (fp)\Big|_{-\pi}^{\pi} = f(\pi)p(\pi) - f(-\pi)p(-\pi) = 0.$$

Lemma 3.1.5. *Let P be a probability measure on $\mathbb{S}^1$ with Lebesgue density p and Stein class $(\mathcal{F}(P), \mathcal{T}_p)$. For all $f \in \mathcal{F}(P)$, $\mathbb{E}_P[\mathcal{T}_p f] = 0$.*

Proof. This statement is evident by Definition 3.1.3. For $f \in \mathcal{F}(P)$ and $\mathcal{I} = \{x \in \mathbb{S}^1 : p(x) > 0\}$,

$$\mathbb{E}_P[\mathcal{T}_p f] = \int_{\mathcal{I}} \frac{(fp)'}{p} p\, dx + \int_{\mathcal{I}^c} fp\, dx = 0.$$

□

Example 3.1.6. We now give some examples of Stein operators for circular distributions, with $f \in \mathcal{F}(P)$ throughout;

a) Uniform measure $\text{U}(\mathbb{S}^1)$ with $p(x) = (2\pi)^{-1}$:

$$\mathcal{T}_p f(x) = f'(x). \tag{3.6}$$

In this particular instance, the Stein class can be explicitly written down using Definition 3.1.3;

$$\mathcal{F}(P) = \{f \in C(\mathbb{S}^1) : f' \in L^1(dx)\}$$

here f' is the derivative in weak sense.

b) von-Mises $\text{VM}(\mu, \kappa)$ with $p(x) = (2\pi I_0(\kappa))^{-1}\text{exp}(\kappa \cos(x - \mu))$,

$$\mathcal{T}_p f(x) = f'(x) - \kappa \sin(x - \mu) f(x).$$

c) Bingham $\text{Bing}(\mu, \kappa)$ with $p(x) = (2\pi I_0(\kappa/2)\text{exp}(\kappa/2))^{-1}\text{exp}(\kappa \cos^2(x - \mu))$;

$$\mathcal{T}_p f(x) = f'(x) - \kappa \sin(2(x - \mu)) f(x).$$

d) Cardiod $\mathrm{C}(\mu, \rho)$ with $p(x) = (2\pi)^{-1}(1 + 2\rho \cos(x - \mu))$ and $|\rho| \leq \frac{1}{2}$;

$$\mathcal{T}_p f(x) = f'(x) - \frac{2\rho \sin(x - \mu)}{2\rho \cos(x - \mu) + 1} f(x).$$

We shall now go about describing a standardized coordinate system for the purposes of integration.

Definition 3.1.7. Let $X \sim P$ be a circular random variable. The mean angle μ is defined as $\mu := \mathrm{Arg}(\mathbb{E}_P[e^{iX}])$ for $i^2 = -1$ and Arg is the complex argument (principal value) function.

Note that this definition of moment exists on $\mathbb{S}^1$, and more importantly, is unique.

Remark. The parameter μ in the above examples b), c), and d) are all the mean angle of their respective distributions.

This particular quantity's origin is from the first circular moment $\mathbb{E}_P[e^{iX}]$ which is decomposed into $\phi_1 = \rho e^{i\mu}$ in what we shall call the standard coordinate system of $\mathbb{S}^1$. That is, when $\mathbb{S}^1$ is viewed as the wrapping of the interval $[-\pi, \pi)$. In ϕ_1, ρ is the mean resultant length and μ is known as the mean direction [MJ09]. Before calculating μ, it is key to determine what coordinate system one is working with. The first moment, ϕ_1, is not necessarily invariant to choice of coordinate system and Arg is only defined to have support in the standard coordinates of $\mathbb{S}^1$. One will have to convert to standard coordinates before proceeding to calculate μ. Under the standard coordinates, one particular property μ has is that $\mathbb{E}_P[\sin(X - \mu)] = 0$. This fact will be important in the section following.

We shall be using the parameter μ as a foundation from which we shall construct a coordinate system on $\mathbb{S}^1$ for the purpose of integration. It will allow us to define and compute integrals in an interval whose midpoint is not necessarily 0. This procedure is as follows: For any $x \in \mathbb{S}^1$ which is not $\mu + \pi$, the antipodal point to μ, there is a unique tangent vector $V_x \in T_\mu \mathbb{S}^1 \cong \mathbb{R}$ with $\sqrt{g(V_x, V_x)} < \pi$

such that $\exp_\mu(V_x) = x$. Therefore, the map $x \mapsto V_x$ determines a local coordinate system covering $\mathbb{S}^1 \setminus \{\mu+\pi\}$. Furthermore, the mapping identifies $\mathbb{S}^1 \setminus \{\mu+\pi\}$ with $[-\pi, \pi) \subset \mathbb{R}$. Under this new coordinate system, μ is identified with the origin of $\mathbb{R}$ and $x - \mu$ is simply V_x. Then, by mapping $\mu + \pi$ to π, we in effect identify $\mathbb{S}^1$ with $(-\pi, \pi] \subset \mathbb{R}$ with the understanding that the two endpoints are wrapped together; $-\pi$ is identified with π. In the case where μ is not unique, for example with the uniform measure on $\mathbb{S}^1$, we take (any) one of the valid values for μ and form the corresponding identification as described above. Hence our chosen coordinate system of $\mathbb{S}^1$ depends on P. Any reference to the μ *coordinate system* will directly refer to this construction. Under the μ *coordinate system*, since we have identified μ with 0, $\mathbb{E}_P[\sin(X)] = 0$ for a random variable X on $\mathbb{S}^1$. Moreover, p is now centred at μ. For example, if P is the von-Mises, $P \sim \mathrm{VM}(\mu, \kappa)$ in standard coordinates, P in the μ *coordinate system* changes to $P \sim \mathrm{VM}(0, \kappa)$.

3.1.2 The Inverse Operator

The next objective is to define the inverse of the Stein operator (3.5) so that we can define the Stein kernel. Under the μ *coordinate system*, since we have identified μ with 0, $\mathbb{E}_P[\sin(X)] = 0$ for a random variable X on $\mathbb{S}^1$.

Definition 3.1.8. Let $\mathcal{F}^0(P) = \{h \in L^1(P) : \mathbb{E}_P[h] = 0\}$ and define the operator $\mathcal{T}_p^{-1} : \mathcal{F}^0(P) \to \mathcal{F}(P)$ by

$$\mathcal{T}_p^{-1}h(x) := \begin{cases} \dfrac{1}{p(x)} \int_{-\pi}^{x} h(y)p(y)dy + \dfrac{h(-\pi)p(-\pi)}{p(x)} & \text{if } p(x) \neq 0, \\ h(x) & \text{if } p(x) = 0, \end{cases} \tag{3.7}$$

where parameters of p are defined in terms of the μ *coordinate system*

We also have the following equality

$$\mathcal{T}_p^{-1}h(x) = -\frac{1}{p(x)} \int_{x}^{\pi} h(y)p(y)dy + \frac{h(\pi)p(\pi)}{p(x)} \quad p(x) \neq 0,$$

which is due to the fact that, for $h \in \mathcal{F}^0(P)$,

$$\mathbb{E}_P[h] = \int_{-\pi}^{x} h(y)p(y)dy + \int_{x}^{\pi} h(y)p(y)dy = 0.$$

In this construction, the integration is performed under the μ *coordinate system.*

Remark. Typically, when working with distributions on $\mathbb{R}$ with support $\mathcal{I} = [a, b]$, one does not see the second term on the right of (3.7). This is due to the fact that one chooses p such that at the end points of the support, $p(a) = p(b) = 0$ and so the extra constant disappears.

Proposition 3.1.9. *$\mathcal{T}_p^{-1}$ is the inverse of $\mathcal{T}_p$.*

Proof. There are two sections to this proof: the first being the case where $p(x) > 0\ \forall x \in \mathbb{S}^1$ and the second being the case where $p(x) \geq 0,\ p(x) = 0$ for some $x \in \mathbb{S}^1$. We begin with the first case. First, let us check that for a function $h \in \mathcal{F}^0(P)$, we have $\mathcal{T}_p\mathcal{T}_p^{-1}h = h$:

$$\begin{aligned}
\mathcal{T}(\mathcal{T}_p^{-1}h)(x) &= \frac{((\mathcal{T}_p^{-1}h(x))p(x))'}{p(x)} \\
&= \frac{1}{p(x)}\frac{\partial}{\partial x}\bigg(\int_{-\pi}^{x} h(y)p(y)dy + h(-\pi)p(-\pi)\bigg) \\
&= \frac{1}{p(x)}h(x)p(x) \\
&= h(x).
\end{aligned}$$

Now to show the other way, let $h \in \mathcal{F}(P)$. Since $\mathbb{E}_P[\mathcal{T}_p h] = 0$ by Lemma 3.1.5, it is clear that $\mathcal{T}_p h \in \mathcal{F}^0(P)$. Then,

$$\begin{aligned}
\mathcal{T}_p^{-1}(\mathcal{T}_p h)(x) &= \frac{1}{p(x)}\int_{-\pi}^{x} \mathcal{T}_p h(y)p(y)dy + \frac{h(-\pi)p(\mu-\pi)}{p(x)} \\
&= \frac{1}{p(x)}\int_{-\pi}^{x} \frac{((h(y)p(y))'}{p(y)}p(y)dy + \frac{h(-\pi)p(-\pi)}{p(x)} \\
&= \frac{1}{p(x)}\int_{-\pi}^{x} (h(y)p(y))'dy + \frac{h(-\pi)p(-\pi)}{p(x)} \\
&= \frac{1}{p(x)}\big(h(x)p(x) - h(-\pi)p(-\pi) + h(-\pi)p(-\pi)\big)
\end{aligned}$$

$$= h(x).$$

For the second case define $\mathcal{I} = \{x \in \mathbb{S}^1 : p(x) > 0\}$ and let $h \in \mathcal{F}^0(P)$. For $x \in \mathcal{I}^c$,

$$\mathcal{T}_p(\mathcal{T}_p^{-1}h)(x) = \mathcal{T}_p h(x) = h(x).$$

For the other way, since $p(x) = 0$ it is clear that $\mathbb{E}_P[\mathbb{I}_{\mathcal{I}^c}\mathcal{T}_p h] = 0$ and hence $\mathcal{T}_p h \in \mathcal{F}^0(P)$ on $\mathcal{I}^c$, therefore

$$\mathcal{T}_p^{-1}(\mathcal{T}_p h)(x) = \mathcal{T}_p^{-1}h(x) = h(x).$$

□

The primary role of the constant $h(-\pi)p(-\pi)$ is to preserve the value $h(-\pi)$ and also to ensure that $\mathcal{T}_p^{-1}$ is injective. If we do not include it, then $\mathcal{T}_p^{-1}(\mathcal{T}_p h)(-\pi) = 0$ and $\mathcal{T}_p^{-1}(\mathcal{T}_p h)(\pi) = 0$ which is not necessarily true. It is now clear that this extra term plays a pivotal role in ensuring that $\mathcal{T}_p^{-1}$ is the inverse of $\mathcal{T}_p$. However, owing to the definition of the Stein operator, we also have the following result:

Corollary 3.1.10. *Fix any $C \in \mathbb{R}$. For $g(x) = \mathcal{T}_p^{-1}h(x) + C/p(x)$ for x such that $p(x) \neq 0$ and $g(x) = h(x)$ if $p(x) = 0$,*

$$\mathcal{T}_p g(x) = \mathcal{T}_p(\mathcal{T}_p^{-1}h)(x).$$

A special quantity is obtained when we select $h(x) = \nu - x$ in (3.7) where $\nu = \int_{-\pi}^{\pi} xp(x)dx$. Applying the inverse operator to this particular h we generate an object known as the classical Stein kernel: for $p(x) \neq 0$

$$\begin{aligned}\tau(x) :&= \mathcal{T}_p^{-1}(\nu - \mathrm{Id})(x)\\ &= \frac{1}{p(x)}\int_{-\pi}^{x}(\nu - y)p(y)dy + \frac{(\nu+\pi)p(-\pi)}{p(x)}\\ &= -\frac{1}{p(x)}\int_{x}^{\pi}(\nu - y)p(y) + \frac{(\nu+\pi)p(\pi)}{p(x)}.\end{aligned} \tag{3.8}$$

Again, we must define $\tau(x) = \nu - x$ when $p(x) = 0$. However, this does follow from Definition 3.1.8.

Example 3.1.11. Let $X \sim \mathrm{U}(\mathbb{S}^1)$, the uniform measure on $\mathbb{S}^1$. Then $\nu = \int_{-\pi}^{\pi} x/2\pi dx = 0$ and choose $\mu = 0$. The Stein kernel of this distribution is

$$\begin{aligned}\tau(x) &= 2\pi \int_{-\pi}^{x} -\frac{y}{2\pi} dy + \pi \\ &= -\frac{y^2}{2}\bigg|_{-\pi}^{x} + \pi \\ &= \frac{\pi^2 - x^2}{2} + \pi. \end{aligned} \tag{3.9}$$

Definition 3.1.12. Let P be a probability distribution on $\mathbb{S}^1$ with $\nu = \int_{-\pi}^{\pi} xp(x)dx$, and $X \sim P$. A Stein kernel of P is the random variable $\tau(X)$ such that

$$\mathbb{E}[\tau(X)\phi'(X)] = \mathbb{E}[(X-\nu)\phi(X)] \tag{3.10}$$

for all differentiable ϕ for which the expectations exist.

From this definition, we can see that the Stein kernel previously defined in Equation (3.8) obeys (3.10): Let $\phi \in C^1(\mathbb{S}^1)$, then

$$\begin{aligned}\mathbb{E}[\tau(X)\phi'(X)] &= \int_{\mathbb{S}^1} \frac{1}{p(x)}\bigg(\int_{-\pi}^{x} (\nu - y)p(y)dy + (\nu+\pi)p(\mu+\pi)\bigg)\phi'(x)p(x)dx \\ &= \int_{\mathbb{S}^1} \phi'(x)\bigg(\int_{-\pi}^{x} (\nu - y)p(y)dy + (\nu+\pi)p(\pi)\bigg)dx \\ &= \phi(x)\bigg(\int_{-\pi}^{x} (\nu - y)p(y)dy + (\nu+\pi)p(\pi)\bigg)\bigg|_{-\pi}^{\pi} + \int_{\mathbb{S}^1} \phi(x)(x-\nu)p(x)dx \\ &= \mathbb{E}[(X-\nu)\phi(X)],\end{aligned}$$

since $\int_{-\pi}^{\pi}(\nu - y)p(y)dy = 0$ and $\phi(-\pi) = \phi(\pi)$.

Other such examples of closed form expressions for the kernel can be seen in [LRS17b, Example 4.9] where Ley et. al. formulate a Stein kernel for the family of Pearson distributions on $\mathbb{R}$.

One of the main uses of the Stein kernel is to be able to construct bounds on the Wasserstein distance between distributions on $\mathbb{R}$ (see [LRS17a, Theorem 3.1]). If we wish to adapt this theorem onto $\mathbb{S}^1$ for a circular distribution, say a von-Mises distribution, we will have to compute the Stein kernel. So far we have only looked at examples with the uniform measure on $\mathbb{S}^1$ due to its simple Lebesgue density. However, for the von-Mises distribution in particular, one will quickly find that obtaining a closed form solution of the kernel is not straightforward, and in some cases impossible (when $\kappa \neq 0$). For simplicity, let $X \sim \mathrm{VM}(0, \kappa)$; then

$$\tau(x) = \mathcal{T}_p(\mathbb{E}[X] - \mathrm{Id})(x) = e^{-\kappa \cos(x)} \int_{-\pi}^{x} (\nu - y) e^{\kappa \cos(y)} dy + \frac{C}{p(x)}.$$

There are two problems with this: The first is that the integral is intractable. We can obtain bounds on $\tau(x)$, but these bounds do not particularly aid in bounding the Wasserstein metric as they will be large — this is akin to bounding the solution to the Stein equation which is what we wanted to avoid. The second is the definition of ν. In Example 3.1.11 we used $\nu = \int_{-\pi}^{\pi} xp(x)dx$, but for directional data analysis this is not used as a parameter of location since the standard mean is not well defined on $\mathbb{S}^1$.

A different approach is to redefine μ to be the intrinsic mean of $\mathbb{S}^1$: $\mathbb{E}[e^{iX}]$. This does, however, require us to completely redefine the kernel to ensure that $\tau(\pi) = 0$. This is precisely the route that we shall take in the next section except we shall not use the full extrinsic mean, rather the mean angle as defined in Definition 3.1.7. We shall see that this particular choice of parameter works well with the von-Mises and other common directional distributions.

3.2 The Circular Stein Kernel

The end of the previous section lead us to motivate the need to redefine the Stein kernel for certain circular distributions. We dedicate this section to constructing this new kernel as well as computing it for a handful of distributions. Similarly to the classical Stein kernel, we shall utilise the inverse Stein operator to initially define it.

Definition 3.2.1. Let X be a circular random variable with distribution P, mean angle $\mu = \mathrm{Arg}(\mathbb{E}[e^{iX}])$ and μ-centred Lebesgue density p; then $\sin(x) \in \mathcal{F}^0(P)$ on the μ *coordinate system*. The *circular Stein kernel* τ^c of P is defined as

$$\begin{aligned}\tau^c(x) &:= \mathcal{T}^{-1}\sin(-\mathrm{Id})(x)\\ &= -\frac{1}{p(x)}\int_{-\pi}^{x}\sin(y)p(y)dy\\ &= \frac{1}{p(x)}\int_{x}^{\pi}\sin(y)p(y)dy.\end{aligned}$$

By a μ-centred density of a circular random variable $X \sim P$ with density $p(x;\mu)$, we mean $p(x;0)$. In Example 3.2.2, we shall see that this μ-centred density is precisely the density of the random variable $X - \mu \bmod 2\pi$.

We are distinguishing the circular Stein kernel from the classical Stein kernel (3.8) with superscript c.

Example 3.2.2. Let $X \sim \mathrm{VM}(\mu,\kappa)$ with Lebesgue density given by $p(x) = (2\pi I_0(\kappa))^{-1}\exp(\kappa\cos(x-\mu))$ with $\mu \in \mathbb{S}^1$ and $\kappa > 0$. To calculate the mean angle, we recall the special function $I_1(\kappa) = \frac{1}{2\pi}\int_0^{2\pi}\cos(x)e^{\kappa\cos(x)}dx$. It turns out that discounting the moment by $e^{i\mu}$ aids in its calculation:

$$\begin{aligned}\mathbb{E}[e^{i(X-\mu)}] &= \frac{1}{2\pi I_0(\kappa)}\int_{\mathbb{S}^1} e^{i(x-\mu)}e^{\kappa\cos(x-\mu)}dx\\ &= \frac{1}{I_0(\kappa)}\bigg(\frac{1}{2\pi}\int_{\mathbb{S}^1}\cos(x-\mu)e^{\kappa\cos(x-\mu)}dx\\ &\qquad\qquad + \frac{i}{2\pi}\int_{\mathbb{S}^1}\sin(x-\mu)e^{\kappa\cos(x-\mu)}dx\bigg)\\ &= \frac{I_1(\kappa)}{I_0(\kappa)},\end{aligned}$$

since $\sin(x)e^{\kappa\cos(x)}$ is antisymmetric about the origin. Therefore, because $\mathrm{Arg}(\mathbb{E}[e^{i(X-\mu)}]) = 0$, it must be that $\mu = \mathrm{Arg}(\mathbb{E}[e^{iX}])$. We then calculate the circular Stein kernel by switching to μ *coordinates*,

$$\begin{aligned}\tau^c(x) &= \exp(-\kappa\cos(x))\int_{-\pi}^{x} -\sin(y)\exp(\kappa\cos(y))dy \\ &= \frac{\exp(-\kappa\cos(x))}{\kappa}\Big(\exp(\kappa\cos(x)) - \exp(\kappa\cos(-\pi))\Big) \\ &= \frac{1}{\kappa} - \frac{1}{\kappa}\exp\big(-\kappa(1+\cos(x))\big).\end{aligned}$$

Notably, we have the following bounds on τ^c:

$$0 \leq \tau^c(x) \leq \frac{1}{\kappa}(1-e^{-2\kappa}) \leq \frac{1}{\kappa} \tag{3.11}$$

where the minimum $\tau^c(x) = 0$ is achieved at $x = \pm\pi$ and the maximum $\tau^c(x) = \frac{1}{\kappa}(1-e^{-2\kappa})$ is achieved at $x = 0$. This particular bound on τ^c for the von-Mises distribution will be of use to us later on.

Example 3.2.3. Let X be a one-dimensional Bingham random variable, with Lebesgue density

$$p(x) = \frac{1}{2\pi e^{\kappa/2} I_0(\frac{\kappa}{2})}\exp\big(\kappa\cos^2(x-\mu)\big), \quad x \in \mathbb{S}^1.$$

One can deduce that the mean angle is μ due to the fact that p is symmetric about μ, and so in standard coordinates $\mathbb{E}[\sin(X-\mu)] = 0$. To calculate the circular Stein kernel of this random variable, we must first compute the integral

$$\begin{aligned}-\int_{-\pi}^{x}\sin(y)\exp\big(\kappa\cos^2(y)\big)dy &= \int_{-\sqrt{\kappa}}^{\sqrt{\kappa}\cos(x)}\frac{e^{z^2}}{\sqrt{\kappa}}dz \\ &= \frac{\sqrt{\pi}}{2\sqrt{\kappa}}\big(\mathrm{erfi}(\sqrt{\kappa}\cos(x)) - \mathrm{erfi}(-\sqrt{\kappa})\big) \\ &= \frac{\sqrt{\pi}}{2\sqrt{\kappa}}\big(\mathrm{erfi}(\sqrt{\kappa}\cos(x)) + \mathrm{erfi}(\sqrt{\kappa})\big),\end{aligned}$$

where, erfi(x) is the imaginary error function,

$$\mathrm{erfi}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{t^2}dt,$$

and relates to the error function $\operatorname{erf}(x) = i\operatorname{erfi}(ix)$ with $i^2 = -1$. Therefore, the circular Stein kernel is

$$\tau^c(x) = \frac{\sqrt{\pi}}{2}\frac{e^{-\kappa\cos^2(x)}}{\sqrt{\kappa}}\big(\operatorname{erfi}(\sqrt{\kappa}\cos(x)) + \operatorname{erfi}(\sqrt{\kappa})\big).$$

Example 3.2.4. Let $X \sim \mathrm{U}(\mathbb{S}^1)$ be the uniform measure on $\mathbb{S}^1$ with Lebesgue density $p(x) = \frac{1}{2\pi}$, $x \in \mathbb{S}^1$, and choose $\mu = 0$ as is done in Example 3.1.11. Then the circular Stein kernel is

$$\tau^c(x) = 2\pi\int_{-\pi}^{x} -\frac{\sin(y)}{2\pi}dy = \cos(x) + 1.$$

One can also obtain this kernel by taking the limit as $\kappa \to 0$ in Example 3.2.2 with the von-Mises distribution for $\mu = 0$;

$$\begin{aligned}\lim_{\kappa\to 0}\frac{1 - e^{\kappa(-1-\cos x)}}{\kappa} &\overset{\text{L'hôp}}{=} \lim_{\kappa\to 0}(1 + \cos x)e^{\kappa(-1-\cos x)}\\ &= 1 + \cos x.\end{aligned}$$

This is possible due to the fact that the Lebesgue density for the von-Mises distribution is continuous in κ, and its limit is the Lebesgue density for the uniform measure.

Example 3.2.5. Let $X \sim \mathrm{C}(\mu, \rho)$, the cardioid distribution. This particular distribution is a slight perturbation of the uniform distribution on $\mathbb{S}^1$. It has Lebesgue density

$$p(x) = \frac{1}{2\pi}(1 + 2\rho\cos(x - \mu)), \quad x \in \mathbb{S}^1,$$

where $\mu \in \mathbb{S}^1$, $|\rho| \leq \frac{1}{2}$. Then $\mathbb{E}[e^{iX}] = \rho e^{i\mu}$ and so the mean angle is μ. Then its circular Stein kernel is

$$\tau^c(x) = -\frac{1}{1 + 2\rho\cos(x)}\int_{-\pi}^{x}\sin(y)(1 + 2\rho\cos(y))dy$$

$$\begin{aligned}
&= -\frac{1}{1+2\rho\cos(x)}\int_{-\pi}^{x}\sin(y)+\rho\sin(2y)dy\\
&= \frac{1}{1+2\rho\cos(x)}\left(\cos(y)+\frac{\rho}{2}\cos(2y)\Big|_{-\pi}^{x}\right)\\
&= \frac{1}{1+2\rho\cos(x)}\left(\frac{\rho}{2}\cos(2x)+\cos(x)+1-\frac{\rho}{2}\right)\\
&= \frac{1+\cos(x)-\rho\sin^2(x)}{1+2\rho\cos(x)}.
\end{aligned}$$

One interesting observation is that the Lebesgue density for a cardioid distribution is in fact a scaled-normalised version of the circular Stein kernel for the uniform distribution.

Similarly to the classical Stein kernel, the circular Stein kernel also satisfies the following integration by parts property.

Lemma 3.2.6. *Define X to be a random variable on $\mathbb{S}^1$ with corresponding circular Stein kernel τ^c and mean angle μ, and let ϕ be absolutely continuous with weak derivative ϕ'. We have that*

$$\mathbb{E}[\sin(X)\phi(X)] = \mathbb{E}[\tau^c(X)\phi'(X)].$$

Proof. Let X have Lebesgue density p on $\mathbb{S}^1$, then

$$\mathbb{E}[\tau^c(X)\phi'(X)] = -\int_{\mathbb{S}^1}\int_{-\pi}^{x}\sin(y)p(y)dy\phi'(x)dx.$$

Using integration by parts with $u = \int_{-\pi}^{x}\sin(y)p(y)dy$ and $v' = \phi'(x)$ we obtain

$$\begin{aligned}
\mathbb{E}[\tau^c(X)\phi'(X)] &= -\phi(x)\int_{-\pi}^{x}\sin(y)p(y)dy\Big|_{x=-\pi}^{\pi} + \int_{\mathbb{S}^1}\sin(x)\phi(x)p(x)dx\\
&= \mathbb{E}[\sin(X)\phi(X)].
\end{aligned}$$

In the second equality we have used the continuity of ϕ and the fact that $\mathbb{E}_P[\sin(X)] = 0$ in the μ *coordinate system.* □

3.3 Bounding of the Wasserstein Distance

Let $W = \{h : \|h'\|_\infty \leq 1\}$ be the set of Lipschitz continuous functions with Lipschitz constant less than or equal to 1. The Wasserstein distance between two probability measures P_1 and P_2 on measurable space $(\Omega, \mathcal{F})$ is defined as

$$d_W(P_1, P_2) = \sup_{h \in W} |\mathbb{E}_{P_1}[h] - \mathbb{E}_{P_2}[h]|.$$

Using the Stein operator, we can construct the Stein equation for $X \sim p$ as

$$\mathcal{T}_p f_h(x) = h(x) - \mathbb{E}[h(X)], \tag{3.12}$$

with $\mathcal{T}_p$ defined in Definition 3.1.4. Clearly, $\mathbb{E}[\mathcal{T}_p f_h(X)] = 0$ since $\mathbb{E}[h(X) - h(X)] = 0$. More concretely, we can say that $h - \mathbb{E}[h(X)] \in \mathcal{F}^0(P)$. It is now evident that we can apply the inverse Stein operator to both sides of the Stein equation (3.12) in order to find its solution. However, by Corollary 3.1.10 we may choose $C = -p(-\pi)(h(-\pi) - \mathbb{E}[h(X)])$ so that we can define the solution

$$f_h(x) := \frac{1}{p(x)} \int_{-\pi}^{x} (h(y) - \mathbb{E}[h(X)])p(y)dy. \tag{3.13}$$

3.3.1 Main Theorem

We next turn our attention to the use of the Stein kernel to bound the Wasserstein distance for distributions on $\mathbb{S}^1$. We shall take a similar approach to that of [LRS17a] with modifications of the kernel that are discussed in [Döb15] since this does not involve bounding the solution to the Stein equation directly.

Lemma 3.3.1. *Let τ^c be the circular Stein kernel of a circular random variable X with Lebesgue density p and mean angle μ. Define the solution to the Stein equation f_h by (3.13) and further define $g_h = f_h/\tau^c$. Then, we have for any*

Lipschitz continuous test function $h : \mathbb{S}^1 \to \mathbb{R}$

$$|g_h(x)| \leq \|h'\|_\infty \frac{\int_{\mu-\pi}^{x} (\mathbb{E}[X] - y)p(y)dy}{\left|\int_{\mu-\pi}^{x} \sin(\mu - y)p(y)dy\right|}.$$

Remark. This result was formulated by Döbler in [Döb15, Proposition 3.13 a)]. Particularly in Döbler's work, he looked at a general kernel on an interval of $\mathbb{R}$ with closure $\overline{(a,b)}$. This kernel took the form

$$\eta(x) = \frac{1}{p(x)} \int_a^x \gamma(t)p(t)dt$$

where γ is a function so that $\eta(b) = 0$. In the proposition, Döbler imposed conditions on the γ that can be used; in particular, γ is decreasing on $\overline{(a,b)}$. However, this condition is not necessary for part a) of the relevant proposition and instead relies upon properties of h and the CDF of p. Therefore this lemma is easily translated from an interval onto the circle.

Remark. Under the μ *coordinate system* we choose $\mathbb{E}[X] = 0$ and $\int_{\mu-\pi}^{x} yp(y)dy$ is viewed purely as an integral and not as an expectation.

Theorem 3.3.2. *Let X and Y be circular random variables with Lebesgue densities p_1, p_2 respectively and* $\operatorname{supp}(X) = \operatorname{supp}(Y) = \mathbb{S}^1$*, define $\pi_0(x) = \frac{p_2(x)}{p_1(x)}$. Furthermore, let μ be the mean angle of X and τ^c be the circular Stein kernel of X. Assume that p_1, p_2 and π_0 are differentiable everywhere on $\mathbb{S}^1$. Then we have the following bounds on the Wasserstein distance between X and Y:*

$$|\mathbb{E}[\tau^c(X)\pi_0'(X)]| \leq d_W(Y, X) \leq \mathbb{E}[|\alpha(X)\pi_0'(X)\tau^c(X)|],$$

where

$$\alpha(x) = \frac{\int_{\mu-\pi}^{x} (\mathbb{E}[X] - y)p_1(y)dy}{\int_{\mu-\pi}^{x} \sin(\mu - y)p_1(y)dy}.$$

Proof. We begin by proving the lower bound.

First, note that since the sine function is Lipschitz continuous with a Lipschitz constant of 1,

$$|\mathbb{E}[\sin(Y)] - \mathbb{E}[\sin(X)]| \leq d_W(Y, X).$$

Moreover, since μ is the mean angle of X, the second expectation on the left hand side is 0. For the first expectation,

$$\begin{aligned}
\mathbb{E}[\sin(Y)] &= \int_{\mathbb{S}^1} \sin(x) p_2(x) dx \\
&= \int_{\mathbb{S}^1} \sin(x) \frac{p_2(x)}{p_1(x)} p_1(x) dx \\
&= \mathbb{E}[\sin(X)\pi_0(X)].
\end{aligned}$$

Then by applying Lemma 3.2.6 with $\phi = \pi_0$, we obtain the lower bound.

For the upper bound, let $(\mathcal{F}_1, \mathcal{T}_1)$ and $(\mathcal{F}_2, \mathcal{T}_2)$ be the Stein pairs of X and Y respectively. Then by the definition of the Stein equation, one clearly sees that $f_h := \mathcal{T}_1^{-1}(h - \mathbb{E}[h(X)]) \in \mathcal{F}_1$ since $h - \mathbb{E}[h(X)] \in \mathcal{F}_1^0$. We need to verify that $f_h \in \mathcal{F}_2$: First $f' \in L^1(dx)$ and f_h is differentiable everywhere on $\mathbb{S}^1$ already, because $f_h \in \mathcal{F}_1$. Furthermore $\int_{\mathbb{S}^1}(f_h p_2)' dx = f_h p_2|_{-\pi}^{\pi} = 0$ by continuity. Whence, we can conclude that $f_h \in \mathcal{F}_2$ and, more importantly, $f_h \in \mathcal{F}_1 \cap \mathcal{F}_2$. Using this fact, we wish to relate the Stein operators of X and Y; $\mathcal{T}_2(f_h)(x) = f_h'(x) + \frac{p_2'(x)}{p_2(x)} f_h(x)$ and $\mathcal{T}_1(f_h)(x) = f_h'(x) + \frac{p_1'(x)}{p_1(x)} f_h(x)$. One can see that both operators share a common term of $f_h'(x)$, and so

$$\mathcal{T}_2(f_h) - \mathcal{T}_2(f_h) = (\log \pi_0)' f_h. \tag{3.14}$$

Now, by definition of the Stein equation (3.13),

$$\begin{aligned}
\mathbb{E}[h(Y)] - \mathbb{E}[h(X)] &= \mathbb{E}[\mathcal{T}_1(f_h)(Y)] \\
&= \mathbb{E}[\mathcal{T}_1(f_h)(Y)] - \mathbb{E}[\mathcal{T}_2(f_h)(Y)] \\
&= -\mathbb{E}[f_h(Y)(\log \pi_0)'(Y))] \\
&= -\mathbb{E}\left[\tau^c(Y) \frac{f_h(Y)}{\tau^c(Y)} (\log \pi_0)'(Y)\right]. \qquad (3.15)
\end{aligned}$$

The second equality is due to the fact that $\mathbb{E}[\mathcal{T}_2(f_h)(Y)] = 0$, since $f_h \in \mathcal{F}_1 \cap \mathcal{F}_2$, and in the third equality we have used Equation (3.14). Define the quantity $g_h = f_h/\tau^c$.

Now, using Lemma 3.3.1,

$$|g_h(x)| \leq \|h'\|_\infty |\alpha(x)|. \tag{3.16}$$

Combining (3.15) and (3.16) together we obtain the upper bound

$$\begin{aligned} d_W(Y, X) &\leq \sup_{h:\|h'\|_\infty \leq 1} \|h'\|_\infty \mathbb{E}[|\tau^c(Y)\alpha(Y)(\log \pi_0)'(Y)|] \\ &= \mathbb{E}[|\tau(Y)\alpha(Y)(\log \pi_0)'(Y)|]. \end{aligned}$$

□

Remark. For probability densities on $\mathbb{R}$, p_1 and p_2 must obey the following requirements of [LRS17a, Theorem 3.1]:

$$\left(\pi_0(x)\int_{-\pi}^{x}(h(y) - \mathbb{E}[h(X_1)])p_1(y)dy\right)' \in L^1(P_2)$$

and

$$\lim_{x\to\pi} \pi_0(x)\int_{-\pi}^{x}(h(y) - \mathbb{E}[h(X_1)])p_1(y)dy = 0.$$

These two assumptions are not essential when looking at $\mathbb{S}^1$. This is due to the compactness of the circle and the continuity of functions at $-\pi$ and π.

It is worth mentioning that $\alpha(x)$ is not a bounded function — it has singularities at $x = \pm\pi$. To tackle this problem, we will be multiplying it by an auxiliary function that comes about as a result of $(\log p)'$. For example, in both von-Mises and Bingham cases, $(\log p)'$ will contain a sine function to assist in removing the singularity.

Lemma 3.3.3. *Let X be a random variable on $\mathbb{S}^1$ with Lebesgue density p such that $p(x) \neq 0$ on $\mathbb{S}^1$. Suppose, without loss of generality, that $\mathbb{E}[X] = 0$ in the*

Euclidean sense after making use of an appropriate chart. Then the function $|\sin(x)\alpha(x)| \leq 2\pi$ *and attains this maximum at* $x = \pm 2\pi$.

Proof. We begin by bounding the function from above.

$$\begin{aligned}\lim_{x\to-\pi}\alpha(x)\sin(x) &= \lim_{x\to-\pi}\frac{\sin(x)\int_{-\pi}^{x}-yp(y)dy}{\int_{-\pi}^{x}-\sin(y)p(y)dy}\\ &= \lim_{x\to-\pi}\frac{\cos(x)\int_{-\pi}^{x}-yp(y)dy - x\sin(x)p(x)}{-\sin(x)p(x)}.\end{aligned}$$

When looking at the absolute value of the function,

$$\begin{aligned}\lim_{x\to-\pi}|\alpha(x)\sin(x)| &= \lim_{x\to-\pi}\frac{|\cos(x)\int_{-\pi}^{x}-yp(y)dy - x\sin(x)p(x)|}{|\sin(x)p(x)|}\\ &\leq \lim_{x\to-\pi}\frac{|\cos(x)\int_{-\pi}^{x}-yp(y)dy|}{|\sin(x)p(x)|}+\pi\\ &= \lim_{x\to-\pi}\frac{-\cos(x)}{p(x)}\lim_{x\to-\pi}\frac{|\int_{-\pi}^{x}-yp(y)dy|}{|\sin(x)|}+\pi\\ &= \lim_{x\to-\pi}\frac{xp(x)}{\cos(x)}\lim_{x\to-\pi}\frac{1}{p(x)}+\pi\\ &= 2\pi.\end{aligned}$$

To show that this is indeed a maximum, first note that the function $\alpha(x)\sin(x)$ satisfies $\alpha(0)\sin(0) = 0$. Denote $m(x) = \int_{-\pi}^{x}-yp(y)dy$ and $s(x) = \int_{-\pi}^{x}-\sin(y)p(y)dy$, then

$$(\alpha(x)\sin(x))' = \frac{1}{s(x)}\left(x\sin(x)p(x) - \frac{m(x)}{s(x)}\sin(x)p(x) + \frac{m(x)}{s(x)}\cos(x)\right)$$

which is 0 if and only if

$$s(x)x\sin(x)p(x) - m(x)\sin(x)p(x) + \cos(x)m(x) = 0.$$

This only occurs at the point $x = \pm\pi$, and because $\alpha(-\pi)\sin(-\pi) > 0$, it attains a maximum. □

Note that the assumption of $\mathbb{E}[X] = 0$ is satisfied if we transformed to μ-

coordinates, since the function $f(x) = x$ is anti-symmetric about the origin.

3.3.2 Applications of Theorem 3.3.2

To end the chapter, we present a handful of examples of using Theorem 3.3.2 to bound the Wasserstein metric and show convergence in distribution.

In the first example, we shall compare two Bayesian posterior densities with the same likelihood, but different priors. This is an analogous to the example in [LRS17a, Section 4.2] which discusses the influence of priors on a normal model of data. In contrast, we explore the effect a von-Mises prior has on a von-Mises model and compare it to a model with uninformative prior. Inference of this type was first performed in [MEA76] which looked at the von-Mises Fisher distribution and classes of priors that can be applied to give analytic results. More recently, a more relevant inference has been performed in [DW99] which specifically uses a von-Mises model (on $\mathbb{S}^1$) with von-Mises prior. This type of inference has been used in finding the location of an airplane locator transmitters [GL88]. This example will be have a base von-Mises model, and we wish to compare a uniform prior with a von-Mises prior.

Example 3.3.4. Let $X_1, ..., X_n$ be iid samples from a VM(μ, κ) distribution. For the purposes of this example, we shall keep κ fixed and will be performing inference on the mean angle μ. The likelihood of μ is calculated to be

$$\begin{aligned} L(\mu; x) &\propto \prod_{i=1}^{n} \exp(\kappa \cos(x_i - \mu)) \\ &= \exp\big(\kappa R \cos(\mu - \psi)\big), \end{aligned} \tag{3.17}$$

with

$$R^2 = n^2(\bar{C}^2 + \bar{S}^2),\ \tan\psi = \frac{\bar{S}}{\bar{C}},$$

where

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} \cos(x_i),\ \bar{S} = \frac{1}{n}\sum_{i=1}^{n} \sin(x_i).$$

Note that the form of the likelihood (3.17) is that of a von-Mises with location and precision parameters ψ and κR respectively.

With the likelihood set up, we will select two priors on μ. The first prior will be the uniform prior on $\mathbb{S}^1$; $\pi_1(\mu) = \frac{1}{2\pi}$. The other shall be an independent (of the data) von-Mises random variable; $\mu \sim \mathrm{VM}(0, \kappa^*)$. We shall name these models, Model 1 and Model 2 respectively.

For Model 1, the posterior density is $\mu|X_1, ..., X_n \sim \mathrm{VM}(\psi, \kappa R)$. For Model 2, the posterior density will change. Similarly to the derivation of the likelihood above, one can check that the posterior of μ under Model 2 is

$$\begin{aligned}\pi_2(\mu|x) &\propto \exp\big(\kappa R\cos(\mu-\psi)\big)\exp\big(\kappa^*\cos\mu\big)\\ &= \exp\big(R'\cos(\mu-\psi')\big),\end{aligned}$$

with

$$R'^2 = \kappa^2 R^2 + (\kappa^*)^2 + 2\kappa\kappa^* R\cos\psi,\ \tan\psi' = \frac{\kappa R\sin\psi}{\kappa R\cos\psi + \kappa^*}.$$

In other words, $\mu|X_1, ..., X_n \sim \mathrm{VM}(\psi', R')$ under Model 2.

Now we can apply Theorem 3.3.2 with $X \sim \mathrm{VM}(\psi, \kappa R)$ and $Y \sim \mathrm{VM}(\psi', R')$. The Radon-Nikodym derivative between these two measures both exists and is differentiable everywhere on $\mathbb{S}^1$; and $\pi_0(x) = \frac{p_2(x)}{p_1(x)} \propto \exp(\kappa^*\cos\mu)$. Using inequality (3.11) we know that, in our case, $\tau^c(x) \leq 1/(\kappa R)$. Then using the fact that $\sup_{x\in\mathbb{S}^1} |\alpha(x)\sin(x)| = 2\pi$, the bound on the Wasserstein metric is

$$\frac{I_1(R')}{I_0(R')}\sin(\psi-\psi') \leq d_W(X,Y) \leq \frac{2\kappa^*\pi}{\kappa n\sqrt{\bar{C}^2+\bar{S}^2}}.$$

This shows a convergence rate of $O(n^{-1})$ for the Wasserstein metric between model 1 and model 2 since $\sqrt{\bar{C}^2+\bar{S}^2}$ is $O(1)$ in probability. In other words, the effect of the prior becomes negligible as n increases to infinity, verifying the Bernstein-von-Mises theorem.

A brief note for the lower bound; as $n \to \infty$ then $\tan'\psi \to \tan\psi$ meaning that

$\sin(\psi - \psi') \to 0$ also.

As previously mentioned, this example is reminiscent of [LRS17a, Section 4.2] where Ley et. al. discuss comparing Bayesian posteriors for normal models. Similarly to above, they compare a posterior with uninformative uniform prior against a posterior with normal conjugate prior. In their results they obtain the following bound: for P_1 the model with uninformative prior and P_2 the model with the conjugate prior,

$$\frac{\sigma^2}{n\delta^2 + \sigma^2}|\bar{x} - \mu| \leq d_W(P_1, P_2) \leq \frac{\sigma^2}{n\delta^2 + \sigma^2}|\bar{x} - \mu| + \sqrt{\frac{2}{\pi}}\frac{\sigma^3}{n\delta\sqrt{\delta^2 n + \sigma^2}}$$

where μ and σ^2 are the mean and variance of the normal data respectively and δ^2 is the variance of the conjugate prior. Their results show a convergence of $O(n^{-3/2})$ compared to ours of $O(n^{-1})$. The discrepancy in orders is most likely due to the fact that we are not calculating the expectation $\mathbb{E}[|\alpha(X)\pi_0'(X)\tau^c(X)|]$ and are instead bounding it above by some enveloping function; thus, giving us a less optimal order bound. If it were possible to explicitly calculate $\mathbb{E}[|\alpha(X)\pi_0'(X)\tau^c(X)|]$ or the α function, then a tighter bound is guaranteed and perhaps an upper bound of $O(n^{-3/2})$ will be the result of this.

Example 3.3.5. As in the previous example, let $X_1, ..., X_n$ be iid samples from a VM(μ, κ) distribution. Again, suppose we wish to compare two Bayesian posteriors: one with a von-Mises prior $\mu \sim$ VM$(0, \kappa^*)$, and the other with a Bingham prior $\mu \sim Bing(0, \zeta)$.

Like before, for Model 1, the posterior density is $\mu|X_1, ..., X_n \sim$ VM(ψ', R'). For Model 2 now, we have the posterior density $\pi(\mu|x) \propto \exp(\kappa R\cos(\mu - \psi) + \zeta\cos^2\mu)$. In this case, $\pi_0(x) \propto \exp(\zeta\cos^2\mu - \kappa^*\cos\mu)$. Therefore, using Theorem 3.3.2 and Lemma 3.3.3 with $\tau^c(x) \leq 1/R'$, we have the upper bound on the Wasserstein metric,

$$d_W(X, Y) \leq \frac{2\pi}{R'}(\kappa^* + 2\zeta).$$

Our next example concerns the distance between two ‘wrapped’ distributions.

A wrapping of a distribution of $\mathbb{R}$ is a wrapping of the density onto $\mathbb{S}^1$. Suppose U is a distribution on $\mathbb{R}$, then the wrapping of U is $V = U \bmod 2\pi$. In terms of density functions, they are described using the following equivalences [MJ09]: Let U have Lebesgue density p_U, the Lebesgue density of the wrapping of U, V, is

$$p_V(\theta) = \sum_{k=-\infty}^{\infty} p_U(\theta + 2\pi k), \quad \theta \in [-\pi, \pi).$$

For example, if $U \sim N(0, \sigma^2)$ then the *wrapped normal distribution* V has density

$$p_V(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \sum_{k=-\infty}^{\infty} \exp\bigg(- \frac{(\theta + 2\pi k)^2}{2\sigma^2} \bigg). \tag{3.18}$$

We say V is distributed according to a wrapped normal distribution by $V \sim \mathrm{WN}(\theta, \sigma^2)$.

Example 3.3.6. Let $Z \sim \mathrm{WN}(0, \sigma^2)$ and $X \sim \mathrm{WC}(0, \gamma)$ a wrapped Cauchy distribution that has probability density function

$$p_{WC}(\theta) = \frac{1}{2\pi} \frac{\sinh \gamma}{\cosh \gamma - \cos \theta}, \quad \theta \in \mathbb{S}^1, \gamma > 0,$$

which can be obtained by applying the geometric series formula on to the pdf when written in terms of the characteristic function.

We may alternatively write the pdf of the wrapped normal distribution (3.18) using the Jacobi theta function (see [MJ09, p. 50]),

$$p_{WN}(\theta) = \frac{1}{2\pi} \vartheta_3(\theta, e^{-\sigma^2/2})$$

which in turn, can be rewritten with the Jacobi triple product

$$\begin{aligned} p_{WN}(\theta) &= \frac{1}{2\pi} \prod_{n=1}^{\infty} (1 - e^{-\sigma^2 n})(1 + e^{-\sigma^2 (n-1/2)} e^{i\theta})(1 + e^{-\sigma^2 (n-1/2)} e^{-i\theta}) \\ &= \frac{1}{2\pi} \prod_{n=1}^{\infty} (1 - e^{-\sigma^2 n})(1 + e^{-2\sigma^2 (n-1/2)} + e^{-\sigma^2 (n-1/2)} (e^{i\theta} + e^{-i\theta})) \end{aligned}$$

$$= \frac{1}{2\pi}\prod_{n=1}^{\infty}(1-e^{-\sigma^2 n})(1+e^{-2\sigma^2(n-1/2)}+2\cos\theta e^{-\sigma^2(n-1/2)}).$$

This form is more amenable to taking logarithms, since we can write it as a sum outside of the logarithm,

$$\log p_{WN}(\theta) = -\log 2\pi + \sum_{n=1}^{\infty}\log(1-e^{-\sigma^2 n})+\log(1+e^{-2\sigma^2(n-1/2)}+2\cos\theta e^{-\sigma^2(n-1/2)}).$$

The derivative $(\log p_{WN})'(\theta)$ is now

$$\begin{aligned}(\log p_{WN})'(\theta) &= \sum_{n=1}^{\infty}\frac{-2\sin\theta e^{-\sigma^2(n-1/2)}}{1+e^{-2\sigma^2(n-1/2)}+2\cos\theta e^{-\sigma^2(n-1/2)}}\\ &= \sum_{n=1}^{\infty}\frac{-\sin\theta}{\cosh(\sigma^2(n-\frac{1}{2}))+\cos\theta}. \end{aligned} \tag{3.19}$$

We shall be using the wrapped Cauchy as our basis for the comparison; it has circular Stein kernel equal to

$$\tau^c(\theta) = (\cosh\gamma-\cos\theta)\log\left(\frac{\cosh\gamma-\cos\theta}{\cosh\gamma+1}\right)$$

and its log derivative is

$$(\log p_{WC})'(\theta) = -\frac{\sin\theta}{\cosh\gamma-\cos\theta}. \tag{3.20}$$

Since this derivative contains a sine function, we may apply Theorem 3.3.2 and Lemma 3.3.3 together with the log derivatives (3.19) and (3.20) to obtain a bound on the Wasserstein metric

$$\begin{aligned} d_W(Z,X) \leq 2\pi\,\mathbb{E}\Bigg[\Bigg|&(\cosh\gamma-\cos X)\log\left(\frac{\cosh\gamma+1}{\cosh\gamma-\cos X}\right)\\ &\times\left(\frac{1}{\cosh\gamma-\cos X}+\sum_{n=1}^{\infty}\frac{1}{\cosh(\sigma^2(n-\frac{1}{2}))+\cos X}\right)\Bigg|\Bigg]. \end{aligned}$$

This expectation is intractable due to the contribution of the wrapped normal term, however it is possible to bound this sum from above:

$$\begin{aligned}
\sup_{\theta\in\mathbb{S}^1}\sum_{n=1}^{\infty}\frac{1}{\cosh(\sigma^2(n-\frac{1}{2}))+\cos(\theta)} &= \sum_{n=1}^{\infty}\frac{1}{\cosh(\frac{\sigma^2}{2}(2n-1))-1}\\
&= \sum_{n\in\mathbb{N}\setminus 2\mathbb{N}}\frac{1}{\cosh(\frac{\sigma^2}{2}n)-1}\\
&\le \sum_{n\in\mathbb{N}\setminus 2\mathbb{N}}\frac{1}{\frac{1}{2}(\frac{\sigma^2}{2}n)^2}\\
&= \frac{8}{\sigma^4}\sum_{n\in\mathbb{N}\setminus 2\mathbb{N}}\frac{1}{n^2}\\
&= \frac{\pi^2}{\sigma^4},
\end{aligned}$$

where in the final equality we have used the zeta function for odd indices. Applying this bound, as well as bounding $\frac{1}{\cosh\gamma-\cos x}$ above by $\frac{1}{\cosh\gamma-1}$, the Wasserstein metric is bounded above by the following expectation

$$d_W(Z,X)\le 2\pi\mathbb{E}\left[(\cosh\gamma-\cos X)\log\left(\frac{\cosh\gamma+1}{\cosh\gamma-\cos X}\right)\right]\left(\frac{1}{\cosh\gamma-1}+\frac{\pi^2}{\sigma^4}\right),$$

in which the expectation can be further evaluated, giving

$$d_W(Z,X)\le 4\pi\sinh\gamma\log(1+e^{-\gamma})\left(\frac{1}{\cosh\gamma-1}+\frac{\pi^2}{\sigma^4}\right). \tag{3.21}$$

By setting $\gamma=\sigma$ and letting $\gamma\to\infty$, d_W converges to 0 with at least leading order $O(\gamma^{-1})$.

Despite using Theorem 3.3.2 as a tool to construct analytic bounds on the Wasserstein metric, we may also estimate this upper bound numerically. In certain cases, like the ones we shall present shortly, numerically integrating the expectation will provide a much more powerful and meaningful bound with which one can use to show that a distribution is sufficiently close to another.

Take for example a comparison between the von-Mises and Bingham distributions. If we were to naïvely apply the methods as above for this comparison, one will eventually end up with a bound

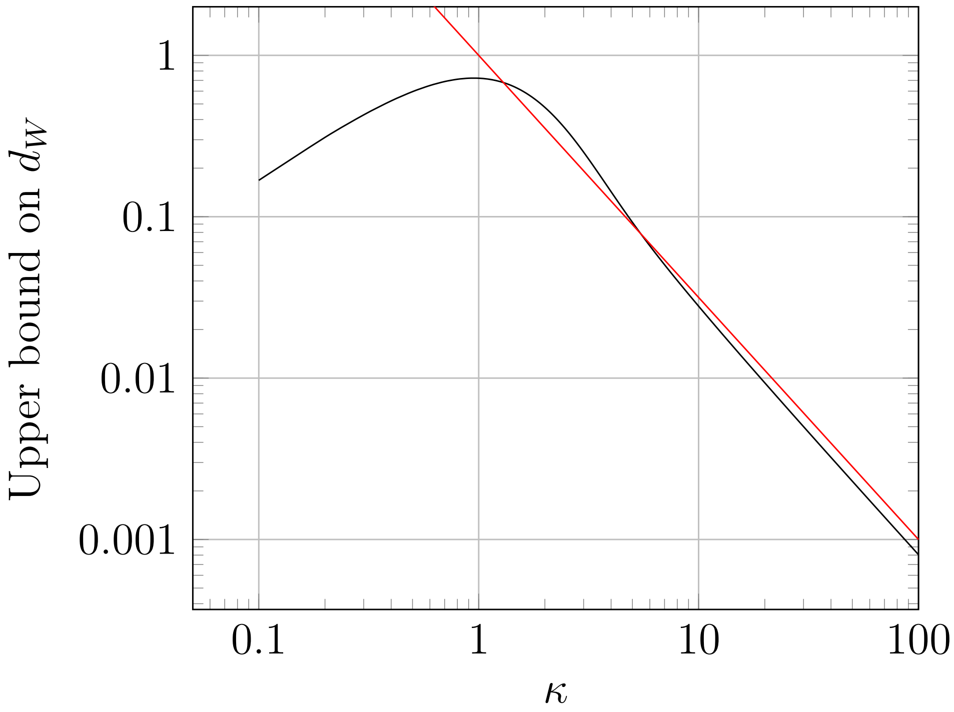

Figure 3.1: A log-log plot of the upper bound on d_W in the case of $\kappa = \zeta$ in black and a plot of the function $\kappa^{-3/2}$ in red.

Suppose $X \sim \mathrm{VM}(0, \kappa)$ and $Y \sim Bing(0, \zeta)$ the Bingham distribution discussed in Example 3.2.3. The Radon-Nikodym derivative between these two distributions is $\pi_0 = \exp(\zeta \cos^2 x - \kappa \cos x)$ and its derivative is easily computed $(\log \pi_0)' = -\sin(x)(2\zeta \cos(x) - \kappa)$ and so using Theorem 3.3.2 and Lemma 3.3.3

$$\begin{aligned} d_W(X, Y) &\leq \frac{1}{\kappa}\mathbb{E}[|\alpha(X)\sin(X)(2\zeta\cos(X) - \kappa)] \\ &\leq 4\pi\frac{\zeta}{\kappa}\mathbb{E}[|\cos(X)|] + 2\pi \end{aligned}$$

This is clearly an extremely uninformative bound on d_W. Convergence of the bound does not go to 0, but to 6π by setting $\kappa = \zeta$ and letting $\kappa \to \infty$. This tells us that 'nice' analytic bounds are only available in some very specific cases. Implementing a numerical integration scheme for the von-Mises and Bingham comparison can be performed quite simply, using Theorem 3.3.2 we approach this as a double integral. Below show the results for the von-Mises and Bingham comparison:

Figure 3.1 shows, approximately, that the leading order of convergence for the von-Mises and Bingham distributions is $\kappa^{-3/2}$ as κ increases. As we might expect, the distributions converge to each other as their respective parameters tend to infinity. The value at which the distance is bounded above by 0.1 occurs approximately when $\kappa = \zeta = 5$. Co-incidence of the distributions at $\kappa = \zeta = 0$ is also adhered to.

The upper bound on the Wasserstein metric can be written as

$$d_W(X,Y) \leq \frac{C}{\kappa^{3/2}} + \frac{c}{\kappa^2}$$

for large κ where C and c are constants and where $X \sim \mathrm{VM}(0,\kappa)$, $Y \sim Bing(0,\kappa)$. The constant C can be numerically approximated with $C = 0.83$ such that all values of the integral calculated are less than $0.83\kappa^{-3/2}$ for each $\kappa > 25$. Thus, for large κ, it is sensible to suggest that

$$d_W(X,Y) \leq \frac{0.83}{\kappa^{3/2}} + \frac{c}{\kappa^2}. \tag{3.22}$$

An example central to directional statistics is the comparison between the von-Mises and Wrapped Normal distributions. It is known (cf. [MJ09, p.38]) that

$$p_{\mathrm{VM}}(\theta;\mu,\kappa) - p_{WN}(\theta;\mu,A(\kappa)) = O(\kappa^{-1/2}), \quad \kappa \to \infty$$

where $A(\kappa) = \frac{I_1(\kappa)}{I_0(\kappa)}$. Likewise with the von-Mises and Bingham comparison, we shall apply a numerical integration scheme in order to calculate the upper bound to the Wasserstein metric.

Figure 3.2 suggests an order of convergence of $O(\kappa^{-1/2})$ for $\sigma^2 = A(\kappa)$, but also for $\sigma^2 = \kappa$. In fact, for smaller κ, our bound on the Wasserstein metric is also lower by using $\sigma^2 = \kappa$.

A similar bound to (3.22) can also be numerically approximated and we con-

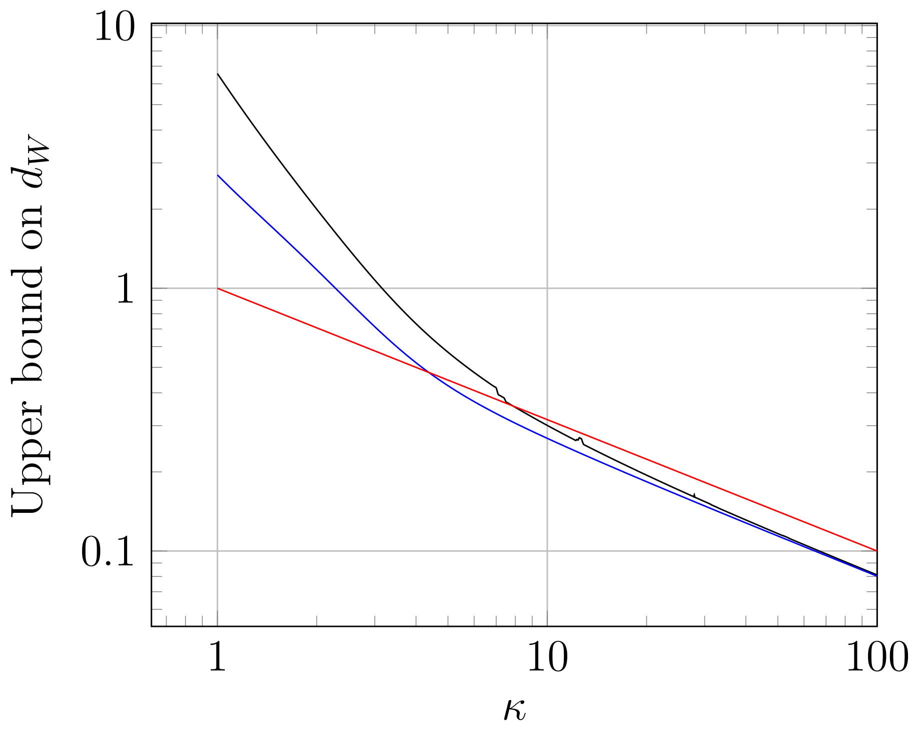

Figure 3.2: A log-log plot of the upper bound on d_W in the case of $\sigma^2 = \frac{I_1(\kappa)}{I_0(\kappa)}$ in black and $\sigma^2 = \kappa$ in blue, and a plot of the function $\kappa^{-1/2}$ in red.

jecture that

$$d_W(X, Z) \leq \frac{0.87}{\kappa^{1/2}} + \frac{c}{\kappa^2}$$

for $\kappa > 50$, where $Z \sim \text{WN}(0, \frac{I_1(\kappa)}{I_0(\kappa)})$ and c is a constant.

Surprisingly, and perhaps coincidentally, for $\kappa > 100$, in both the von-Mises and Wrapped normal and von-Mises and Bingham comparisons, the constant C coincided at 0.804.

3.4 Conclusion

To conclude the chapter, we summarise our findings and present some potential topics that require further research.

The essence of this chapter has been modifying the existing framework of the density approach for Stein's method to cater for the geometry of the circle. This was motivated by the fact that the Bakry-Èmery-Ricci criterion was not satisfied for many popular distributions in directional statistics. As such we could not use the diffusion approach detailed in [LLBF22] and we instead pursued a density approach.

It was found that a handful of conditions could be relaxed when transferring from intervals of $\mathbb{R}$ to $\mathbb{S}^1$. Particularly, the fact that for $f, p \in C^0(\mathbb{S}^1)$, $f(-\pi)p(-\pi) = f(\pi)p(\pi)$. Moreover, we could relax the restriction of fp being absolutely continuous to being in $L^1(\mathbb{S}^1)$.

For the purposes of integration, a standard chart on the circle had to be selected. This led us to defining the mean angle μ in Definition 3.1.7 which we used as our 'zero point' around which we centred all our probability distributions.

The Stein operator and its inverse were defined in Definitions 3.1.4 and 3.1.8 respectively, and we noted the difference the inverse operator had with the typical inverse operator for intervals on $\mathbb{R}$. With these in hand, it was necessary to redefine the Stein kernel in Definition 3.2.1, which included the newly defined mean angle μ. With these tools in hand, Theorem 3.3.2 was formulated which gave us a bound on the Wasserstein metric between distributions on $\mathbb{S}^1$. When compared to its Euclidean counterpart from [LRS17a, Theorem 3.1] it was found that further assumptions were automatically satisfied by taking into account the geometry of the circle.

The chapter culminated in Section 3.3.2 where multiple examples were explored. Two of which were the comparison between two Bayesian models and another between a wrapped normal distribution and a wrapped Cauchy distribution. We also numerically approximated the upper bound to the Wasserstein metric in two cases. The most important being the comparison between the wrapped normal and the von-Mises distribution. We found that the upper bound had a leading order of $\kappa^{-1/2}$ which verified previous asymptotic results.

One observation that was noted, however, was the fact that if we wanted an analytic, closed form, bound on the Wasserstein metric, then it was imperative that our quotient $\pi_0(x)$ contained a $\sin(x)$ so that we could apply Lemma 3.3.3.

This is a key weakness of the upper bound of Theorem 3.3.2.

With regards to the numerical approximation of the upper bound in Theorem 3.3.2, it would be interesting to explore the comparison the approximation of the upper bound of the Wasserstein metric and the approximation of the Wasserstein metric using computational optimal transport, to determine how good the upper bound in Theorem 3.3.2 is — if such a method can indeed be implemented.

In the next chapter, we shall be departing with the density approach in favour of the diffusion approach to construct brand new bounds on the Wasserstein metric on more interesting manifolds like $\mathbb{H}^n$ and $\mathrm{SO}(n)$.

Chapter 4

Wasserstein Bounds on Manifolds

This chapter will be dedicated to constructing analytical bounds of the Wasserstein metric for random variables X and Z with prescribed distributions on specific manifolds — abuses of notation shall happen often such as $X \sim p$ meaning that X is distributed with density function p. The manifolds in question are all complete, simply connected Riemannian manifolds. We shall be looking at: comparisons between the uniform and von-Mises-Fisher distributions on $\mathbb{S}^n$, a comparison between the Hyperbolic heat kernel and Riemannian-Gaussian distributions on $\mathbb{H}^3$, comparisons between Matrix von-Mises-Fisher distributions on $\mathrm{SO}(n)$, and finally a comparison on the space of symmetric positive definite matrices $\mathcal{P}_n$. The manifolds we examine are popular for data analysis, particularly the von-Mises Fisher distribution on $\mathbb{S}^n$. As discussed in Section 1.1.3, the general procedure of constructing a bound on the Wasserstein metric between two probability measures $d\mu_\phi \propto e^{-\phi} d\mathrm{vol}$ and $d\mu_\psi \propto e^{-\psi} d\mathrm{vol}$ via the diffusion approach requires that each distribution satisfies the Bakry-Émery-Ricci criterion

$$\mathrm{Ric} + \mathrm{Hess}^\phi \geq 2\kappa g \tag{4.1}$$

for some $\kappa > 0$.

We restate Theorem 1.1.7 as it will be our primary tool for each comparison we make.

Theorem 4.0.1 ([LLBF22]). *Suppose $X \sim \mu_\phi$ and $Z \sim \mu_\psi$ are two probability distributions on a manifold M. If two such distributions satisfy the Bakry-Émery-Ricci criterion* (4.1) *for the same $\kappa > 0$, an upper bound on the Wassestein metric is*

$$d_W(X, Z) \leq \frac{1}{2\kappa}\mathbb{E}[|\nabla(\psi - \phi)(X)|]. \tag{4.2}$$

Proof. Given that each distribution satisfies the Bakry-Émery-Ricci criterion, we use the diffusion approach to construct a Stein method individually for each distribution. Let $\mathcal{A}_1 = \frac{1}{2}\Delta_M - \frac{1}{2}g(\nabla\phi, \nabla)$ and $\mathcal{A}_2 = \frac{1}{2}\Delta_M - \frac{1}{2}g(\nabla\psi, \nabla)$ be the Stein operators for X and Z respectively.

We can subtract $\mathcal{A}_2$ from $\mathcal{A}_1$ to obtain a new operator

$$L := \mathcal{A}_1 - \mathcal{A}_2 = \frac{1}{2}g(\nabla(\psi - \phi), \nabla).$$

We draw attention to the fact that under X, $\mathbb{E}_{\mu_\phi}[Lf(X)] = -\mathbb{E}_{\mu_\phi}[\mathcal{A}_2 f(X)]$ since $\mathcal{A}_1$ is a Stein operator of μ_ϕ. Now, writing the Stein equation out and taking expectations with respect to μ_ψ we have

$$\begin{aligned} \mathbb{E}_{\mu_\phi}[h(X)] - \mathbb{E}_{\mu_\psi}[h(Z)] &= \mathbb{E}_{\mu_\phi}[\mathcal{A}_2 f_h(X)] \\ &= -\mathbb{E}_{\mu_\phi}[L f_h(X)]. \end{aligned}$$

Taking norms and supremum over the space of functions with Lipschitz constant less than or equal to 1 gives the results. □

In order to show that distributions on $\mathbb{S}^n$ and $\mathbb{H}^n$ satisfy the Bakry-Émery-Ricci criterion (4.1), we require some additional results on the Hessian of radially symmetric manifolds.

Suppose that M is an n-dimensional *spherically symmetric* Riemannian manifold (sometimes called a *manifold with pole*). That is, the manifold is prescribed with a metric of the form $g = d\rho^2 + G(\rho)^2 d\theta^2$ in which ρ is the geodesic distance from a fixed point $o \in M$ called the pole, $d\theta^2$ is the canonical metric of $\mathbb{S}^{n-1}$ and

$G : \mathbb{R}^+ \to \mathbb{R}$ is a continuous function such that $G(0) = 0$ and $G'(0) = 1$.

For $\mathbb{S}^n$, defining the function $G(\rho) = \sin(\rho)$ yields the canonical metric $g = d\rho^2 + \sin^2(\rho)d\theta^2$ on $\mathbb{S}^n$. For $\mathbb{H}^n$, defining the function $G(\rho) = \sinh(\rho)$ yields the canonical metric $g = d\rho^2 + \sinh^2(\rho)d\theta^2$ on the Hyperboloid model of $\mathbb{H}^n$.

When the metric of a Riemannian manifold takes this form, the Hessian of the distance function ρ has an explicit form (cf. [GW06, Proposition 2.20]);

$$\mathrm{Hess}^\rho = \frac{G'(\rho)}{G(\rho)}(g - d\rho \otimes d\rho). \tag{4.3}$$

Additionally, for any function $f \in C^2(\mathbb{R}^+)$,

$$\begin{aligned}\mathrm{Hess}^{f(\rho)} &= D^2 f(\rho),\\ &= f''(\rho)d\rho \otimes d\rho + f'(\rho)D^2\rho,\\ &= f''(\rho)d\rho \otimes d\rho + f'(\rho)\mathrm{Hess}^\rho. \end{aligned}\tag{4.4}$$

These two features of the Hessian of spherically symmetric manifolds shall be used extensively in this chapter to determine whether the Bakry-Émery-Ricci criterion is satisfied.

4.1 Comparison on $\mathbb{S}^n$

The first manifold we shall explore in this chapter will be the n-sphere $\mathbb{S}^n$ in $\mathbb{R}^{n+1}$. We endow the sphere with its canonical Riemannian metric $g = d\rho^2 + \sin^2\rho\, d\theta^2$ where $d\theta^2$ is the metric of $\mathbb{S}^{n-1}$. It is clear that the sphere can be regarded as a spherically symmetric manifold and the function G in this case is $G(\rho) = \sin\rho$. By Equation (4.3), the Hessian of the geodesic distance is

$$\mathrm{Hess}^\rho = \frac{\cos\rho}{\sin\rho}(g - d\rho \otimes d\rho). \tag{4.5}$$

We frequently use the fact that any point $o \in \mathbb{S}^n$ can be regarded as a pole. The volume form on $\mathbb{S}^n$ is $d\text{vol} = \sin^{n-1}\rho\, d\rho d\theta$.

We shall investigate comparisons between three types of distributions on $\mathbb{S}^n$: the uniform, the von-Mises Fisher, and the Fisher Watson. The von-Mises Fisher distribution is the n-dimensional analogue of the von-Mises distribution on $\mathbb{S}^1$ investigated in Chapter 3. In terms of the intrinsic coordinates on $\mathbb{S}^n$, the probability density function (with respect to the volume measure) of the von-Mises Fisher distribution is

$$p_{VMF}(\theta) = \frac{\lambda^{(n-1)/2}}{(2\pi)^{(n+1)/2} I_{(n-1)/2}(\lambda)} e^{\lambda \cos\rho(\theta)}, \quad \theta \in \mathbb{S}^n,\ \lambda > 0, \tag{4.6}$$

where $\rho(\theta) = \rho(\theta, o)$ is the geodesic distance from a point θ on $\mathbb{S}^n$ to a fixed point o. For brevity, we write that the argument of the pdf is $\rho := \rho(\theta)$ since $\mathbb{S}^n$ and p are spherically symmetric about o. For a random variable that is distributed with pdf (4.6), we write $X \sim \text{VMF}(\mu, \lambda)$.

A more commonly used representation is to embed the sphere into $\mathbb{R}^{n+1}$ and use extrinsic coordinates,

$$p_{VMF}(x) = \frac{\lambda^{(n-1)/2}}{(2\pi)^{(n+1)/2} I_{(n-1)/2}(\lambda)} e^{\lambda\langle x, \mu\rangle}, \quad |x| = |\mu| = 1, \kappa > 0 \tag{4.7}$$

where $\langle\cdot,\cdot\rangle$ is the inner product (dot product) in $\mathbb{R}^n$ and where μ is the location parameter equivalent to o in the intrinsic form. We typically abuse this notation and end up writing μ and x for both intrinsic and extrinsic forms. When $n = 1$ we obtain the pdf von-Mises distribution defined earlier in Equation (3.1).

Proposition 4.1.1. *The function $\phi = -\lambda\cos\rho$ satisfies the Bakry-Émery-Ricci criterion* (4.1) *on $M = \mathbb{S}^n$ for $0 < \lambda < n - 1$.*

Proof. We shall apply (4.4) to the function $\phi(\rho) = -\lambda\cos\rho$,

$$\text{Hess}^{\phi} = \lambda\cos\rho\, d\rho \otimes d\rho + \lambda\sin\rho\, \text{Hess}^{\rho}$$

$$
\begin{aligned}
&= \lambda \cos\rho \, d\rho \otimes d\rho + \lambda \sin\rho \frac{\cos\rho}{\sin\rho}(g - d\rho \otimes d\rho) \\
&= \lambda \cos\rho \, g \\
&\geq -\lambda g.
\end{aligned}
$$

The sphere is an Einstein manifold, and hence the Ricci curvature of $\mathbb{S}^n$ is simply a multiple of g; $\mathrm{Ric} = (n-1)g$. Therefore, the von-Mises Fisher distribution satisfies the Bakry-Émery-Ricci criterion if and only if $(n-1) - \lambda > 0$ and so we demand that $0 < \lambda < n-1$. □

Note it is imperative that $n > 1$ for the Bakry-Émery-Ricci criterion to ever be satisfied. Therefore going ahead, we restrict $n > 1$ in all cases.

Remark. The parameter κ is chosen such that the Bakry-Èmery-Ricci criterion is satisfied. In the case of Proposition 4.1.1, we choose $\kappa = \frac{1}{2}(n-1-\lambda)$. Then under the assumption that $n-1 > \lambda > 0$,

$$0 < 2\frac{1}{2}(n-1-\lambda)g \leq \mathrm{Ric} + \mathrm{Hess}^{\phi}.$$

We could always choose κ to be smaller as well. For example, if $n = 5$ and $\lambda = 2$, $\mathrm{Ric} + \mathrm{Hess}^{\phi} \geq 2g \geq g$ and so $\kappa = \frac{1}{2}$ is sufficient in this case. Although, because of the effect of κ in 4.0.1, we would always want to maximize κ so that the Bakry-Èmery-Ricci criterion is satisfied.

Define Z to be the uniform measure on $\mathbb{S}^n$, we write $Z \sim \mathrm{U}(\mathbb{S}^n)$ for brevity. By this we mean, for $Z \sim \mu$, $d\mu = \frac{1}{C_n} d\mathrm{vol}$ where $C_n = \int_{\mathbb{S}^n} d\mathrm{vol} = \mathrm{vol}(\mathbb{S}^n)$. A formula for the volume of the n-sphere is presented later on.

The first comparison we present is between the von-Mises Fisher distribution and uniform measure. It is obvious that the uniform measure on $\mathbb{S}^n$ satisfies the Bakry-Émery-Ricci criterion since the Ricci curvature is positive everywhere, and so we are able to apply Theorem 4.0.1.

Proposition 4.1.2. *Let $X \sim \mathrm{VMF}(o,\lambda)$ and $Z \sim \mathrm{U}(\mathbb{S}^n)$. Then we have the following upper bound on the Wasserstein metric between X and Z:*

$$d_W(Z,X) \leq \frac{\lambda}{2\kappa}\frac{\Gamma\big(\frac{n+1}{2}\big)^2}{\Gamma\big(\frac{n+2}{2}\big)\Gamma\big(\frac{n}{2}\big)}.$$

Proof. Define $\phi = 0$ and $\psi = \lambda\cos\rho$ for the uniform and von-Mises Fisher measures respectively. We obtain via differentiation in ρ alone that $\nabla(\phi-\psi) = \lambda\sin\rho\,\partial_\rho$ and hence $|\nabla(\phi-\psi)| = \lambda\sin\rho$ since $|\partial_\rho| = 1$. The expectation in the upper bound of the Wasserstein metric (4.2) is now

$$\begin{aligned}
\lambda\mathbb{E}_\mu[\sin\rho(Z)] &= \frac{\lambda}{C_n}\int_{\mathbb{S}^n}\sin^n\rho\,d\rho d\theta\\
&= \lambda\bigg(\int_0^\pi \sin^{n-1}\rho\,d\rho\int_{\mathbb{S}^{n-1}}d\theta\bigg)^{-1}\int_0^\pi \sin^n\rho\,d\rho\int_{\mathbb{S}^{n-1}}d\theta\\
&= \lambda\int_0^\pi \sin^n\rho\,d\rho\bigg(\int_0^\pi \sin^{n-1}\rho\,d\rho\bigg)^{-1}\\
&= \lambda\frac{I_n}{I_{n-1}},
\end{aligned}$$

where the integral $I_n = \int_0^\pi \sin^n x\,dx$ can be computed and has the closed form

$$I_n = \sqrt{\pi}\frac{\Gamma\big(\frac{n+1}{2}\big)}{\Gamma\big(\frac{n+2}{2}\big)}. \tag{4.8}$$

Therefore by using (4.8), the ratio I_n/I_{n-1} is simplified to

$$\frac{I_n}{I_{n-1}} = \frac{\Gamma\big(\frac{n+1}{2}\big)^2}{\Gamma\big(\frac{n+2}{2}\big)\Gamma\big(\frac{n}{2}\big)}.$$

The application of Theorem 4.0.1 delivers to us the following bound on the Wasserstein metric:

$$d_W(Z,X) \leq \frac{\lambda}{2\kappa}\frac{\Gamma\big(\frac{n+1}{2}\big)^2}{\Gamma\big(\frac{n+2}{2}\big)\Gamma\big(\frac{n}{2}\big)}$$

for some $\kappa > 0$ when $\lambda < n-1$. □

In the limit as $n \to \infty$, the ratio $\frac{I_n}{I_{n-1}}$ will tend to 1, $\lim_{n\to\infty} d_W(Z,X) \leq \frac{\lambda}{2\kappa}$ which one can show via Stirling's approximation. A further verification that this upper bound makes sense arises from looking at the limit as $\lambda \to 0$. In this case, the probability density of the von-Mises Fisher distribution converges to the uniform density on $\mathbb{S}^n$. When we apply the limit in the bound of the Wasserstein metric it is clear that $\lim_{\lambda\to 0} d_W(Z, X) = 0$ (for small enough λ we can choose $\kappa = \frac{1}{2}$).

The next example demonstrates comparisons between two von-Mises Fisher distributions. Let $X \sim \text{VMF}(o_1, \lambda_1)$ and $Y \sim \text{VMF}(o_2, \lambda_2)$, then the probability densities for both X and Y satisfy the Bakry-Émery-Ricci criterion (4.1) by Proposition 4.1.1.

With X and Y defined above, $\phi(x) = \lambda_1 \cos\rho(o_1, x)$ and $\psi(x) = \lambda_2 \cos\rho(o_2, x)$ respectively. In order to combine these two expressions to obtain $|\nabla(\phi - \psi)|$, we rely upon the extrinsic coordinate form of the density function (4.7). We then write

$$\begin{aligned}
-\phi(x) + \psi(x) &= \lambda_1 \cos\rho(o_1, x) - \lambda_2 \cos\rho(o_1, x) \\
&= \lambda_1 \langle o_1, x\rangle - \lambda_2 \langle o_2, x\rangle \\
&= \langle \lambda_1 o_1 - \lambda_2 o_2, x\rangle \\
&= |\lambda_1 o_1 - \lambda_2 o_2| \left\langle \frac{\lambda_1 o_1 - \lambda_2 o_2}{|\lambda_1 o_1 - \lambda_2 o_2|}, x \right\rangle \\
&= \lambda^* \langle o^*, x\rangle \\
&= \lambda^* \cos\rho(o^*, x)
\end{aligned}$$

where $\lambda^* = |\lambda_1 o_1 - \lambda_2 o_2|$ to ensure that $o^* \in \mathbb{S}^n$. We can then immediately write down the gradient

$$|\nabla(\phi - \psi)(x)| = \lambda^* |\nabla \cos\rho(o^*, x)| = \lambda^* \sin\rho(o^*, x).$$

Whence, by Theorem 4.0.1, an upper bound on the Wasserstein metric is

$$d_W(X, Y) \le \frac{1}{2\kappa} \lambda^* \mathbb{E}_X[\sin\rho(o^*, X)]$$

under the assumption that $0 < \lambda_1, \lambda_2 < n-1$. For general o_1 and o_2 this expectation is intractable and one will have to rely upon Monte Carlo simulations to evaluate it. However, we can obtain an analytic bound in the case where $o_1 = o_2 = o$.

Proposition 4.1.3. *Let $X \sim \mathrm{VMF}(o, \lambda_1)$ and $Y \sim \mathrm{VMF}(o, \lambda_2)$ be von-Mises Fisher random variables on $\mathbb{S}^n$ with $0 < \lambda_1, \lambda_2 < n-1$. Define the function*

$$s_n(\lambda) = \sqrt{\frac{1}{\lambda} \frac{I_{\frac{n}{2}}(\lambda)}{I_{\frac{n-1}{2}}(\lambda)}}.$$

Then the Wasserstein distance between X and Y has an upper bound

$$d_W(Y, X) \leq \frac{|\lambda_1 - \lambda_2|}{\sqrt{2}\kappa} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} s_n(\lambda_1) \wedge s_n(\lambda_2).$$

Proof. We begin by defining $\phi = -\lambda_1 \cos\rho$ and $\psi = -\lambda_2 \cos\rho$. Then, similar to the last example, $|\nabla(\phi - \psi)| = |\lambda_1 - \lambda_2| \sin\rho$. To compute the expectation in Theorem 4.0.1 we must now evaluate the following integral:

$$\int_{\mathbb{S}^n} \sin\rho e^{\lambda_1 \cos\rho} \, d\mathrm{vol} = \mathrm{vol}(\mathbb{S}^{n-1}) \int_0^\pi \sin^n \rho e^{\lambda_1 \cos\rho} d\rho.$$

We see that this new integral is in fact related to the normalizing constant of the von-Mises Fisher distribution on $\mathbb{S}^{n+1}$. If we denote c_n to be the normalizing constant on $\mathbb{S}^n$, then it is known that

$$c_n(\lambda) = \frac{(2\pi)^{(n+1)/2} I_{(n-1)/2}(\lambda)}{\lambda^{(n-1)/2}}.$$

We now evaluate the following integral using c_n;

$$\int_0^\pi \sin^{n-1} \rho e^{\lambda_1 \cos\rho} \, d\rho = \frac{c_n(\lambda_1)}{\mathrm{vol}(\mathbb{S}^{n-1})}$$

for any $n \in \mathbb{N} \setminus \{1\}$. At this point, it is possible to evaluate the expectation:

$$\begin{aligned}\mathbb{E}_X[|\lambda_1 - \lambda_2| \sin\rho] &= |\lambda_1 - \lambda_2| \int_{\mathbb{S}^n} \sin\rho \frac{e^{\lambda_1 \cos\rho}}{c_n(\lambda_1)} d\mathrm{vol} \\ &= |\lambda_1 - \lambda_2| \frac{\mathrm{vol}(\mathbb{S}^{n-1})}{c_n(\lambda_1)} \int_0^\pi \sin^n \rho e^{\lambda_1 \cos\rho}\, d\rho \\ &= |\lambda_1 - \lambda_2| \frac{\mathrm{vol}(\mathbb{S}^{n-1})}{c_n(\lambda_1)} \frac{c_{n+1}(\lambda_1)}{\mathrm{vol}(\mathbb{S}^n)}.\end{aligned}$$

The volume of the $(n-1)$-sphere is known to be

$$\mathrm{vol}(\mathbb{S}^{n-1}) = \frac{2\pi^{n/2}}{\Gamma\left(\frac{n}{2}\right)}.$$

Simplifying the expression for the expectation gives us

$$\begin{aligned}\frac{\mathrm{vol}(\mathbb{S}^{n-1})}{c_n(\lambda)} \frac{c_{n+1}(\lambda)}{\mathrm{vol}(\mathbb{S}^n)} &= \frac{2\pi^{n/2}}{\Gamma\left(\frac{n}{2}\right)} \frac{\Gamma\left(\frac{n+1}{2}\right)}{2\pi^{(n+1)/2}} \frac{(2\pi)^{(n+2)/2} I_{\frac{n}{2}}(\lambda)}{\lambda^{n/2}} \frac{\lambda^{(n-1)/2}}{(2\pi)^{(n+1)/2} I_{\frac{n-1}{2}}(\lambda)} \\ &= \sqrt{\frac{2}{\lambda}} \frac{I_{\frac{n}{2}}(\lambda)}{I_{\frac{n-1}{2}}(\lambda)} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}.\end{aligned}$$

Whence, the expectation required is

$$\mathbb{E}_X[|\lambda_1 - \lambda_2| \sin\rho] = |\lambda_1 - \lambda_2| \sqrt{\frac{2}{\lambda_1}} \frac{I_{\frac{n}{2}}(\lambda_1)}{I_{\frac{n-1}{2}}(\lambda_1)} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}.$$

We can then apply the same approach, but to instead calculate the expectation with respect to Y. This will instead give

$$\mathbb{E}_Y[|\lambda_1 - \lambda_2| \sin\rho] = |\lambda_1 - \lambda_2| \sqrt{\frac{2}{\lambda_2}} \frac{I_{\frac{n}{2}}(\lambda_2)}{I_{\frac{n-1}{2}}(\lambda_2)} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}.$$

Therefore, we may apply Theorem 4.0.1 and the minimum between the two expectations, either taken over X or Y. □

As $n \to \infty$, $s_n(\lambda)\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \to 0$ independently of λ. The comparison between the two distributions is dominated by quantity $\lambda_1 - \lambda_2$, κ can be chosen to be $1/2$ again so long as either λ_1 or λ_2 aren't too big.

We stress that both λ_1 and λ_2 must satisfy the Bakry-Èmery-Ricci criterion for the diffusion approach and Theorem 4.0.1 to be applicable.

The final distribution that we introduce in this section is the Fisher–Watson distribution on $\mathbb{S}^n$. This is a distribution which has a density function of the form

$$p_{FW}(\theta) \propto \exp\bigl(\lambda_1 \cos^2 \rho(\theta, o_1) + \lambda_2 \cos \rho(\theta, o_2)\bigr), \quad \theta, o_1, o_2 \in \mathbb{S}^n,\ \lambda_1, \lambda_2 \geq 0$$

and we have the constraint that $\rho(o_1, o_2) = \frac{\pi}{2}$. We denote by $\mathrm{FW}(o_1, \lambda_1, o_2, \lambda_2)$ the density function above. A useful way to re-express this density function is to use an extrinsic coordinate system on $\mathbb{S}^n$ when viewed in the ambient space of $\mathbb{R}^{n+1}$ like we did with the von-Mises Fisher distribution;

$$p_{FW}(x) \propto \exp\bigl(\lambda_1 \langle x, x_1\rangle^2 + \lambda_2 \langle x, x_2\rangle\bigr), \quad x \in \mathbb{R}^{n+1}, \tag{4.9}$$

where x_1 and x_2 are the extrinsic versions of o_1 and o_2 respectively and satisfy $\langle x_1, x_2\rangle = 0$. Like with all other distributions we have looked at, if we wish to compare this distribution to another, we must verify that it satisfies the Bakry-Émery-Ricci criterion (4.1).

Proposition 4.1.4. *Let $\phi = -\lambda\langle \mu, x\rangle^2$. Then the Hessian of ϕ has the following bound*

$$\mathrm{Hess}^{\phi} \geq -4\lambda g.$$

Proof. Consider the following path $\gamma : [0, 2\pi] \to \mathbb{S}^n$ as $\gamma(t) = x \cos t + v \sin t$ where $\langle x, v\rangle = 0$ are points on $\mathbb{S}^n$ and v is unit speed, then γ is a geodesic on $\mathbb{S}^n$. Recall that the definition of Hess^{ϕ} in terms of curves is

$$\mathrm{Hess}^{\phi} = \left.\frac{d^2\phi(\gamma(t))}{dt^2}\right|_{t=0}.$$

We first begin by simplifying $\phi(\gamma(t))$ before differentiating:

$$\phi(\gamma(t)) = -\lambda\langle \mu, x\cos t + v \sin t\rangle^2$$

$$= -\lambda \cos^2 t \, \langle \mu, x \rangle^2 - \lambda \sin^2 t \, \langle \mu, v \rangle - 2\lambda \sin t \cos t \, \langle \mu, x \rangle \langle \mu, v \rangle.$$

Then the first and second derivatives are

$$\frac{d\phi(\gamma(t))}{dt} = -\lambda(-\sin 2t \, \langle \mu, x \rangle^2 + \sin 2t \, \langle \mu, v \rangle^2 + 2 \cos 2t \, \langle \mu, x \rangle \langle \mu, v \rangle),$$

$$\frac{d^2\phi(\gamma(t))}{dt^2} = -\lambda(-2\cos 2t \, \langle \mu, x \rangle^2 + 2\cos 2t \, \langle \mu, v \rangle^2 - 4 \sin 2t \, \langle \mu, x \rangle \langle \mu, v \rangle).$$

Finally, setting $t = 0$ yields the value of the Hessian and a lower bound

$$\text{Hess}^{\phi} = 2\lambda(\langle \mu, x \rangle^2 - \langle \mu, v \rangle^2) \geq -4\lambda$$

since μ, x and v are all on $\mathbb{S}^n$. □

We now combine Proposition 4.1.1 with Proposition 4.1.4 so that densities of the form (4.9) satisfy the Bakry-Émery-Ricci criterion for $0 < 4\lambda_1 + \lambda_2 < n - 1$.

For the analytical simplicity and tractability, we shall be looking at the case where $\lambda_2 = 0$.

Proposition 4.1.5. *Let $X \sim \mathrm{FW}(x_1, \lambda, x_2, 0)$ and $Z \sim \mathrm{U}(\mathbb{S}^n)$ on $\mathbb{S}^n$. Assume that $0 < \lambda < \frac{n-1}{4}$. Then the Wasserstein distance between X and Z is bounded above;*

$$d_W(X, Z) \leq \frac{1}{\kappa\sqrt{\pi}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \frac{1}{{}_1F_1(\frac{1}{2}, \frac{n+1}{2}, \lambda)} e^{\lambda} \lambda^{-\frac{1}{2}(n-1)} \gamma\Big(\frac{n+1}{2}, \lambda\Big).$$

Here, γ is the lower incomplete Gamma function;

$$\gamma(n, y) = \int_0^y x^{n-1} e^{-x} dx,$$

and ${}_1F_1$ is the Kummer confluent hypergeometric function defined by

$${}_1F_1(a, b, z) = \sum_{n=0}^{\infty} \frac{(a)_n z^n}{(b)_n n!}$$

where $(a)_0 = 1$ and $(a)_n = a(a+1)(a+2)...(a+n-1)$ for $n > 0$.

Proof. Before we begin, the normalizing constant for a Fisher–Watson density with $\lambda_2 = 0$ is

$$\begin{aligned}\int_{\mathbb{S}^n} e^{\lambda\cos^2\rho} d\mathrm{vol} &= \mathrm{vol}(\mathbb{S}^{n-1})\int_0^\pi \sin^{n-1}\rho e^{\lambda\cos^2\rho}\,d\rho \\ &= \sqrt{\pi}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n+1}{2})}\,{}_1F_1\Big(\frac{1}{2},\frac{n+1}{2},\lambda\Big)\mathrm{vol}(S^{n-1}).\end{aligned}$$

For brevity, we shall ignore the normalising constant in the forthcoming calculations and factor it on at the end. We now apply Theorem 4.0.1 with $\phi = -\lambda\cos^2\rho$ and $\psi = 1$:

$$\begin{aligned}\mathbb{E}_X[|\nabla(\phi-\psi)|] &= 2\lambda\mathbb{E}_X[\sin\rho\,|\cos\rho|] \\ &\propto 2\lambda\mathrm{vol}(\mathbb{S}^{n-1})\int_0^\pi \sin^n\rho\,|\cos\rho|\,e^{\lambda\cos^2\rho}\,d\rho \\ &= 2\lambda\mathrm{vol}(\mathbb{S}^{n-1})\Big(\int_0^{\frac{\pi}{2}}\sin^n\rho\,\cos\rho\,e^{\lambda\cos^2\rho}\,d\rho - \int_{\frac{\pi}{2}}^\pi \sin^n\rho\,\cos\rho\,e^{\lambda\cos^2\rho}\,d\rho\Big).\end{aligned}$$

By the anti-symmetry of cosine and symmetry of sine around $\frac{\pi}{2}$, these integrals are negatives of each other and so

$$\mathbb{E}_X[|\nabla(\phi-\psi)|] \propto 4\lambda\mathrm{vol}(\mathbb{S}^{n-1})\int_0^{\frac{\pi}{2}}\sin^n\rho\,\cos\rho\,e^{\lambda\cos^2\rho}\,d\rho.$$

We can simplify this integral by way of the substitution $u = \lambda\sin^2\rho$, then we have

$$\int_0^{\frac{\pi}{2}}\sin^n\rho\,\cos\rho\,e^{\lambda\cos^2\rho}\,d\rho = \frac{1}{2}e^\lambda\lambda^{-\frac{1}{2}(n+1)}\gamma\Big(\frac{n+1}{2},\lambda\Big).$$

Finally, applying Theorem 4.0.1 and accounting for the normalizing constant yields the desired result. □

We may use this result in order to construct an upper bound on the Wasserstein distance between the von-Mises Fisher and Fisher Watson distributions. The idea is that we shall use the fact that d_W is the metric distance of a metric space on probability measures and we shall apply the triangle inequality, with the uniform measure as an auxiliary measure between the two.

4.2 Comparison on $\mathbb{H}^3$

The next manifold that we shall pursue is the 3-dimensional Hyperbolic space $\mathbb{H}^3$. In particular, we shall be studying hyperbolic space using the hyperboloid model, which is an embedding inside of 4-Minowski space together with the Riemannian metric

$$g = d\rho^2 + \sinh^2 \rho \, d\theta^2.$$

Similarly to how it was defined in Section 4.1, ρ denotes the geodesic distance from a fixed point $o \in \mathbb{H}^3$ to $x \in \mathbb{H}^3$, and $d\theta^2$ is the canonical metric of $\mathbb{S}^2$. The volume form of this particular model is $d\text{vol} = \sinh^2 \rho \, d\rho d\theta$. Since $\mathbb{H}^3$ has negative sectional curvature, the cut-locus at every point in $\mathbb{H}^3$ is empty.

The main aim of this section is to compare the two following probability distributions situated on $\mathbb{H}^3$: the probability heat kernel, and the Riemannian-Gaussian distribution. We first introduce the probability heat kernel on $\mathbb{H}^3$. Let $(X_t)_{t\in\mathbb{R}^+}$ be a Brownian motion on $\mathbb{H}^3$. The process has corresponding infinitesimal generator $\mathcal{A} = \frac{1}{2}\Delta_M$, with Δ_M being the Laplace-Beltrami operator on $\mathbb{H}^3$. From Section 2.2, the infinitesimal generator and semigroup satisfy

$$\frac{\partial}{\partial t} P_t f = \mathcal{A} P_t f = P_t \mathcal{A} f, \tag{4.10}$$

then $P_t f = e^{\frac{1}{2}\Delta_M} f$. In terms of an integral, the semigroup can alternatively be defined as $P_t f(x) := \mathbb{E}[f(X_t)|X_0 = x]$, or more explicitly,

$$P_t f(x) = \begin{cases} \int_{\mathbb{H}^3} p(t,x,y) f(y) dy & t > 0, \\ f(x) & t = 0. \end{cases}$$

Since this form of P_t must also satisfy Equation (4.10), by the backward Kolmogorov equation,

$$\frac{\partial}{\partial t}p(t,x,y) = \frac{1}{2}\Delta_M p(t,x,y), \tag{4.11}$$

where the Laplace-Beltrami operator Δ_M acts upon the variable x. For simplicity, we shall fix $y = o \in \mathbb{H}^3$ and instead rewrite the heat kernel as $p(t,x)$.

For a general manifold M, one cannot obtain a closed form solution of (4.11). However for $\mathbb{H}^3$ (and in fact for $\mathbb{H}^{2n+1}$ for $n \in \mathrm{N}$), a closed form solution to the PDE (4.11) exists [GN98]: Let $\rho := \rho(o,x)$, then

$$p_{\mathrm{HK}}(t,\rho) = \frac{1}{(4\pi)^{\frac{3}{2}}}\frac{\rho}{\sinh\rho}e^{-t-\frac{\rho^2}{4t}}, \quad t > 0.$$

For further simplicity, we shall make a time change $k = \frac{1}{4t}$, and drop the dependence on k in the density function so as to treat it like a distributional parameter;

$$p_{\mathrm{HK}}(\rho) = \left(\frac{k}{\pi}\right)^{\frac{3}{2}} e^{-\frac{1}{4k}}\frac{\rho}{\sinh\rho}e^{-k\rho^2}, \quad k > 0. \tag{4.12}$$

The distribution with which we shall compare to the probability heat kernel to is the Riemannian-Gaussian distribution. The name of this distribution is derived from the form of the density function. For $\rho(o,x)$, the geodesic distance from points o to x, the Riemannian-Gaussian distribution on a general manifold M has the probability density function

$$p_{\mathrm{RG}}(\rho) = \frac{1}{C}e^{-c\rho^2}, \quad c > 0 \tag{4.13}$$

where C is the normalizing constant. This is essentially a naïve extension of the normal distribution on $\mathbb{R}^n$ to general manifold. On $\mathbb{R}^n$ it is the case that $\rho(x,y) = |x-y|$ and, moreover, the heat kernel of a Brownian motion on $\mathbb{R}^n$ is described by the pdf (4.13), but this is not necessarily true for general Riemannian manifolds.

In the case of $M = \mathbb{H}^3$, the normalizing constant C is calculated as

$$\begin{aligned} C &= \int_{\mathbb{H}^3} e^{-c\rho^2} d\text{vol} \\ &= \int_{\mathbb{H}^3} e^{-c\rho^2} \sinh^2 \rho \, d\rho d\theta \\ &= \text{vol}(\mathbb{S}^2) \int_0^\infty e^{-c\rho^2} \sinh^2 \rho \, d\rho \\ &= \frac{\pi^{\frac{3}{2}}(e^{\frac{1}{c}} - 1)}{\sqrt{c}}. \end{aligned}$$

A short time asymptotic result of Varadhan in stochastic analysis [Hsu02] tells us that if p_M is the density of the heat kernel on a complete manifold M, then for $x \in M$ and $y \notin \text{Cut}(x)$,

$$\lim_{t \to 0} t \log p_M(t, x, y) = -\frac{1}{2}\rho(x, y)^2. \tag{4.14}$$

Because of this relation, we should expect that as $t \to 0$, our Wasserstein distance between the two distributions under (4.12) and (4.13) also goes to 0.

Consider the random variables $X \sim p_{\text{HK}}$ and $Y \sim p_{\text{RG}}$. To use Theorem 4.0.1, both $\phi := -\log p_{\text{HK}}$ and $\psi := -\log p_{\text{RG}}$ must satisfy the Bakry-Émery-Ricci criterion (4.1). The explicit forms of ϕ and ψ are

$$\phi = -\frac{3}{2}\log k + \frac{3}{2}\pi + \frac{1}{4k} - \log\rho + \log\sinh\rho + k\rho^2, \tag{4.15}$$

$$\psi = c\rho^2 - \log C. \tag{4.16}$$

Proposition 4.2.1. *Both ϕ and ψ in* (4.15) *and* (4.16) *satisfy the Bakry-Émery-Ricci criterion* (4.1) *for $k > 1$ and $c > 1$ respectively.*

Proof. For $M = \mathbb{H}^3$, our function G in equation (4.3) is $G(\rho) = \sinh(\rho)$, and so we can write the Hessian of ρ as

$$\text{Hess}^\rho = \coth\rho\,(g - d\rho \otimes d\rho). \tag{4.17}$$

Now using the relation in (4.4) with $f(\rho) = \rho^2$ together with equation (4.17), we obtain the Hessian of the squared distance:

$$\mathrm{Hess}^{\rho^2} = 2\, d\rho \otimes d\rho + 2\rho \coth \rho\, (g - d\rho \otimes d\rho). \tag{4.18}$$

We start by first proving that ψ satisfies the Bakry-Émery-Ricci criterion for $c > 1$: For ψ, we have, using equation (4.18)

$$\mathrm{Hess}^{\psi} = 2c\, d\rho \otimes d\rho + 2c\rho \coth \rho\, (g - d\rho \otimes d\rho).$$

Now, consider two cases: The first case, let $U \in T_x\mathbb{H}^3$ with $g(\nabla\rho, U) = 0$. Then,

$$\mathrm{Hess}^{\psi}(U, U) = 2c\rho \coth \rho\, g(U, U).$$

Noting that the function $x \coth x \geq 1$ for all $x \geq 0$, we obtain the upper bound

$$\mathrm{Hess}^{\psi}(U, U) \geq 2cg(U, U). \tag{4.19}$$

On the other hand, when $g(\nabla\rho, U) \neq 0$, by the Cauchy-Schwarz inequality, one sees that $g(\nabla\rho, U)^2 \leq g(\nabla\rho, \nabla\rho)g(U, U) = g(\partial_\rho, \partial_\rho)g(U, U) = g(U, U)$. Whence, $g(U, U) - g(\nabla\rho, U)^2 \geq 0$. Therefore, together with the fact that $x \coth x \geq 1$ we have that

$$\mathrm{Hess}^{\psi}(U, U) \geq 2cg(\nabla\rho, U)^2 + 2c(g(U, U) - g(\nabla\rho, U)^2) = 2cg(U, U).$$

Remark. In fact, we have just shown that

$$\mathrm{Hess}^{\rho^2} \geq 2g \tag{4.20}$$

which shall serve us again later on in the proof for computing a lower bound for Hess^{ϕ}.

Since we have obtained a lower bound on the Hessian, we can go on to verify for what values of c the Bakry-Émery-Ricci criterion (4.1) is valid. For $\mathbb{H}^3$, it is well known that $\text{Ric} = -2g$, and so

$$\text{Ric} + \text{Hess}^{\psi} \geq 2cg - 2g = 2(c-1)g$$

which is greater than 0 if and only if $c > 1$.

Now, for ϕ, its first derivative is

$$\nabla\phi = k\nabla\rho^2 + \left(-\frac{1}{\rho} + \coth\rho \right)\nabla\rho$$

and therefore the second derivative is

$$\text{Hess}^{\phi} = k\text{Hess}^{\rho^2} + \left(-\frac{1}{\rho} + \coth\rho \right)\text{Hess}^{\rho} + \left(\frac{1}{\rho^2} - \text{cosech}^2\rho \right) d\rho \otimes d\rho.$$

Since we already have a lower bound on Hess^{ρ^2} we shall forget about this term momentarily and concentrate on the latter two terms. Denote

$$T = \left(-\frac{1}{\rho} + \coth\rho \right)\text{Hess}^{\rho} + \left(\frac{1}{\rho^2} - \text{cosech}^2\rho \right) d\rho \otimes d\rho.$$

Using the same technique as we did with ψ above, we split our tangent space $T_x\mathbb{H}^3$ into two cases: When $U \in T_x\mathbb{H}^3$ with $g(\nabla\rho, U) = 0$, using (4.17)

$$T(U,U) = \left(-\frac{1}{\rho} + \coth\rho \right) \coth\rho g(U,U).$$

Then, noting that the function $\coth^2 x - x^{-1}\coth x \geq 0$ for all $x \geq 0$, $T(U,U) \geq 0$. When $g(\nabla\rho, U) \neq 0$,

$$\begin{aligned} T(U,U) = &\left(-\frac{1}{\rho} + \coth\rho \right) \coth\rho \left(g(U,U) - g(\nabla\rho, U)^2 \right) \\ &+ \left(\frac{1}{\rho^2} - \text{cosech}^2\rho \right) g(\nabla\rho, U)^2 \end{aligned}$$

The first term we bound above by 0 using the fact that $g(U,U) - g(\nabla\rho, U)^2 \geq 0$ and $\coth^2 x - x^{-1}\coth x \geq 0$. The second term we also bound above by 0 because $x^{-2} - \text{cosech}^2 x \geq 0$ for all $x \geq 0$.

To culminate the proof, we have the bound

$$\text{Hess}^{\phi} \geq k\text{Hess}^{\rho^2}.$$

But by equation (4.18),

$$\text{Hess}^{\phi} \geq 2kg$$

and, so noting again that $\text{Ric} = -2g$,

$$\text{Ric} + \text{Hess}^{\phi} \geq 2(k-1)g$$

which is greater than 0 if and only if $k > 1$. □

It is now possible to use Theorem 4.0.1 to formulate a bound between the two distributions.

Proposition 4.2.2. *Let $X \sim p_{\text{HK}}$ and $Y \sim p_{\text{RG}}$ be as described above with respective densities $e^{-\phi}$ and $e^{-\psi}$ see* (4.15) *and* (4.16)*. Then we have the following bound on the Wasserstein metric, for a chosen $\kappa > 0$ that ensures the Bakry-Émery-Ricci criterion is satisfied,*

$$d_W(X,Y) \leq \frac{1}{\kappa}\frac{e^{-\frac{1}{4k}}}{\sqrt{k\pi}}(k + |c-k|) + \frac{2}{\kappa}\left(\frac{|c-k|(2k+1)}{4k} + \frac{1}{4} - \frac{k}{2}\right)\text{erf}\left(\frac{1}{2\sqrt{k}}\right).$$

Proof. We shall utilise Theorem 4.0.1 in order to formulate this bound.

First, we shall go about calculating the quantity $|\nabla(\psi - \phi)|$;

$$\begin{aligned}
|\nabla(\psi-\phi)| &= \left|\left(2(c-k)\rho + \frac{1}{\rho} - \coth\rho\right)\nabla\rho\right| \\
&\leq \left(2|c-k|\rho + \coth\rho - \frac{1}{\rho}\right)|\nabla\rho| \\
&\leq 2|c-k|\rho + \coth\rho - \frac{1}{\rho},
\end{aligned}$$

due to the fact that $x^{-1} - \coth x \leq 0$ for $x \geq 0$ and the fact that $g(\nabla\rho, \nabla\rho) = g(\dot\gamma(0), \dot\gamma(0)) = 1$. The next step is to calculate the expectation:

$$\begin{aligned}
\mathbb{E}[|\nabla(\psi - \phi)(X)|] &= \int_{\mathbb{H}^3} \left(2|c-k|\rho + \coth\rho - \frac{1}{\rho}\right) d\mu_\phi, \\
&= \left(\frac{k}{\pi}\right)^{\frac{3}{2}} e^{-\frac{1}{4k}} \int_{\mathbb{H}^3} \left(2|c-k|\rho + \coth\rho - \frac{1}{\rho}\right) \frac{\rho}{\sinh\rho} \sinh^2\rho \, e^{-k\rho^2} d\rho d\theta, \\
&= 4\pi \left(\frac{k}{\pi}\right)^{\frac{3}{2}} e^{-\frac{1}{4k}} \int_0^\infty \left(2|c-k|\rho + \coth\rho - \frac{1}{\rho}\right) \rho \sinh\rho \, e^{-k\rho^2} d\rho, \\
&= 4k\sqrt{\frac{k}{\pi}} e^{-\frac{1}{4k}} \int_0^\infty 2|c-k|\rho^2 \sinh\rho \, e^{-k\rho^2} + \rho\cosh\rho \, e^{-k\rho^2} - \sinh\rho \, e^{-k\rho^2} d\rho
\end{aligned}$$

to which we split this integral into its three base parts;

$$\begin{aligned}
\int_0^\infty \rho^2 \sinh\rho \, e^{-k\rho^2} d\rho &= \frac{1}{4k^2} + \sqrt{\frac{\pi}{k}} \frac{2k+1}{8k^2} e^{\frac{1}{4k}} \operatorname{erf}\left(\frac{1}{2\sqrt{k}}\right), \\
\int_0^\infty \rho\cosh\rho \, e^{-k\rho^2} d\rho &= \frac{1}{2k} + \frac{1}{4k}\sqrt{\frac{\pi}{k}} e^{\frac{1}{4k}} \operatorname{erf}\left(\frac{1}{2\sqrt{k}}\right), \\
\int_0^\infty \sinh\rho \, e^{-k\rho^2} d\rho &= \frac{1}{2}\sqrt{\frac{\pi}{k}} e^{\frac{1}{4k}} \operatorname{erf}\left(\frac{1}{2\sqrt{k}}\right).
\end{aligned}$$

The next step is to combine the integrals whilst we ignore the normalising constant,

$$\begin{aligned}
&\int_0^\infty \left(2|c-k|\rho + \coth\rho - \frac{1}{\rho}\right) \rho\sinh\rho \, e^{-k\rho^2} d\rho \\
&\quad = \frac{|c-k|}{2k^2} + \frac{|c-k|(2k+1)}{4k^2} \sqrt{\frac{\pi}{k}} e^{\frac{1}{4k}} \operatorname{erf}\left(\frac{1}{2\sqrt{k}}\right) + \frac{1}{2k} \\
&\qquad + \frac{1}{4k}\sqrt{\frac{\pi}{k}} e^{\frac{1}{4k}} \operatorname{erf}\left(\frac{1}{2\sqrt{k}}\right) - \frac{1}{2}\sqrt{\frac{\pi}{k}} e^{\frac{1}{4k}} \operatorname{erf}\left(\frac{1}{2\sqrt{k}}\right) \\
&\quad = \frac{k + |c-k|}{2k^2} + e^{\frac{1}{4k}} \sqrt{\frac{\pi}{k}} \left(\frac{|c-k|(2k+1)}{4k^2} + \frac{1}{4k} - \frac{1}{2}\right) \operatorname{erf}\left(\frac{1}{2\sqrt{k}}\right).
\end{aligned}$$

Then multiplying by the normalising constant,

$$\mathbb{E}[|\nabla(\psi - \phi)(X)|] = \frac{2}{\sqrt{k\pi}} e^{-\frac{1}{4k}} (k + |c-k|) + 4\left(\frac{|c-k|(2k+1)}{4k} + \frac{1}{4} - \frac{k}{2}\right) \operatorname{erf}\left(\frac{1}{2\sqrt{k}}\right).$$

And finally, multiplying by $\frac{1}{2\kappa}$ yields the bound on the Wasserstein metric. □

An interesting consequence of this result comes when enforcing $c = k > 1$.

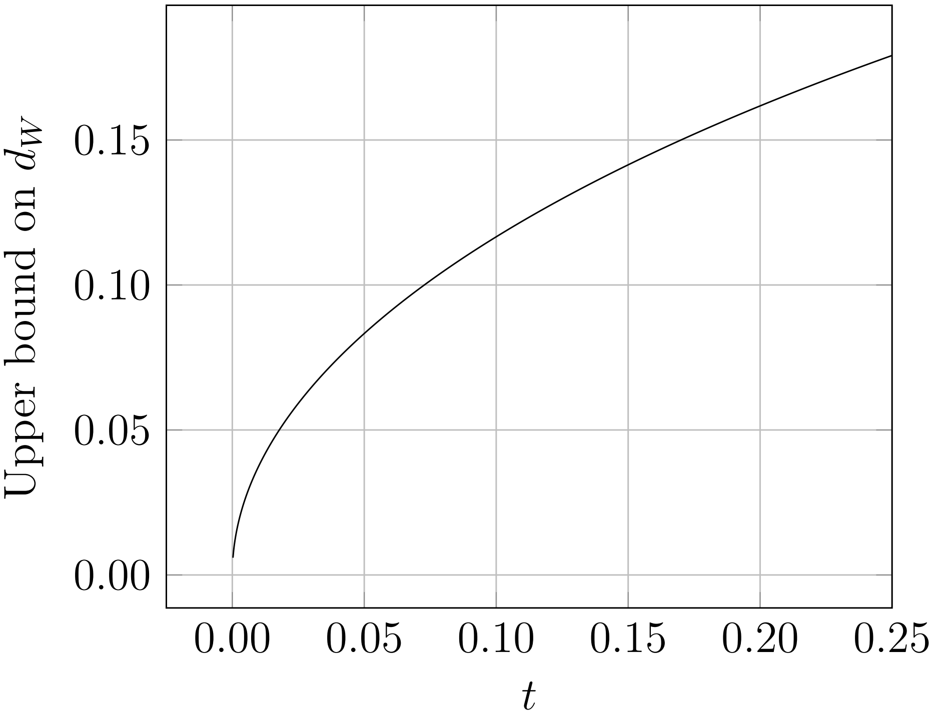

Figure 4.1: Plot of the upper bound on d_W in Proposition 4.2.2 in the case of $c = k$, in terms of $t = 1/4k$.

With this, the bound on the Wasserstein metric reduces to

$$d_W(X,Y) \leq \frac{1}{\kappa}\sqrt{\frac{k}{\pi}}e^{-\frac{1}{4k}} + \frac{2}{\kappa}\left(\frac{1}{4} - \frac{k}{2}\right)\operatorname{erf}\left(\frac{1}{2\sqrt{k}}\right).$$

Then, one may take the limit

$$\lim_{k\to\infty} d_W(X,Y) \leq \lim_{k\to\infty} \frac{1}{\kappa}\sqrt{\frac{k}{\pi}}e^{-\frac{1}{4k}} + \frac{2}{\kappa}\left(\frac{1}{4} - \frac{k}{2}\right)\operatorname{erf}\left(\frac{1}{2\sqrt{k}}\right) = 0,$$

If one back-transforms to $t = \frac{1}{4k}$, then this is akin to letting t tend to 0 see Figure 4.2. In other words, we can infer a short-time asymptotics result $\lim_{t\to 0} d_W(Y,X) = 0$. This result coincides with that of Varadhan's asymptotic relation (4.14) and in fact strengthens it for $\mathbb{H}^3$. For finite $t < \frac{1}{4}$;

$$d_W(X,Y) \leq \frac{1}{2\kappa}\sqrt{\frac{1}{\pi t}}e^{-t} + \frac{2}{\kappa}\left(\frac{1}{4} - \frac{1}{8t}\right)\operatorname{erf}(\sqrt{t}).$$

4.3 Comparison on SO(n)

The penultimate space we shall look at is $\mathrm{SO}(n) = \{S \in \mathbb{R}^{n\times n} : S^\intercal S = SS^\intercal = I_n\}$, the special orthogonal group of dimension n. The Lie algebra of SO(n), $\mathfrak{so}(n) = \{A \in \mathbb{R}^{n\times n} : A^\intercal = -A\}$, is the vector space of $n \times n$ skew-symmetric matrices. The tangent space at the identity $T_I\mathrm{SO}(n) = \mathfrak{so}(n)$. At the point S which is not necessarily the identity, $T_S\mathrm{SO}(n) = \{S\} \times \mathfrak{so}(n)$, and we can pushforward a vector A from $T_S\mathrm{SO}(n)$ to $\mathfrak{so}(n)$ by $S^{-1}A$. We associate SO(n) with the bi-invariant metric $g(E_1, E_2) = -\frac{1}{2}\mathrm{Tr}(E_1E_2)$ for $E_1, E_2 \in \mathfrak{so}(n)$. The Ricci curvature of SO(n) has the form

$$\mathrm{Ric}(E, E) = -\frac{1}{4}B(E, E)$$

where $B(E, E) = \mathrm{Tr}(\mathrm{ad}(E) \circ \mathrm{ad}(E))$ is the Killing form. For SO(n) it is well known that $B(E, E) = (n - 2)\mathrm{Tr}(E^2)$ and so the Ricci curvature is expressed as

$$\mathrm{Ric}(E, E) = \frac{n-2}{2}g(E, E)_I. \tag{4.21}$$

It is worth briefly mentioning that since SO(n) is a closed and compact manifold, two points in SO(n) have two or more geodesics that connect them. For example, on SO(3), we may take one path $\gamma(t) = e^{\pi t K}$ for some $K \in \mathfrak{so}(3)$. Then the endpoints $\gamma(1)$ coincide for K and $K^\intercal$.

For more information on the geometrical theory behind Lie groups, see Appendix B.

One of the most well known distributions on SO(n) is the matrix von-Mises-Fisher distribution, the extension of the von-Mises distribution on $\mathbb{S}^1$ to SO(n). It has density proportional to

$$p_{MVM}(S) \propto \exp\bigl(c\mathrm{Tr}(S_0S)\bigr), \quad c > 0, S_0, S \in \mathrm{SO}(n)$$

and so we set $\phi = -c\mathrm{Tr}(S_0S)$. The parameter S_0 acts as a location-type parameter, whereas c acts as a scaling or precision parameter. When $S_0 = I$ and $n = 2$, we retain the von-Mises distribution on $\mathbb{S}^1$ with $\kappa = 2c$. This density is absolutely continuous with respect to the volume measure on SO(n) and since we are working with a Lie group, this is the Haar measure. It is a measure defined in such a way that it is invariant to left and right translation on the entire Lie group.

In order to make use of Theorem 4.0.1 for SO(n), we have two requirements; the Bakry-Émery-Ricci criterion is satisfied as well as a formula for $\nabla\phi$. In particular, we require the existence of $\nabla\phi$ in order to construct a diffusion whose invariant density is p_{MVM}.

Set $S_0 = I$ to begin with and by define a curve γ on SO(n) which has the properties that $\gamma(0) = S$ and $\dot{\gamma}(0) = SE$ for any $S \in \mathrm{SO}(n)$ and $E \in \mathfrak{so}(n)$. We parametrise the curve via the exponential mapping $\gamma(t) = Se^{tE}$. For any $S \in \mathrm{SO}(n)$, the exterior derivative of ϕ is defined as

$$\begin{aligned} d\phi_S(SE) &= \frac{d}{dt}\phi(\gamma(t))\bigg|_{t=0} \\ &= \lim_{t\to 0}\frac{\phi(Se^{tE}) - \phi(S)}{t} \\ &= \lim_{t\to 0}\frac{-c\mathrm{Tr}(Se^{tE} - S)}{t} \\ &= -c\mathrm{Tr}\left(S\lim_{t\to 0}\frac{e^{tE} - I}{t}\right), \end{aligned}$$

since Tr is a linear operator. Whence,

$$d\phi_S(SE) = -c\mathrm{Tr}(SE).$$

An interesting observation that one finds is that for $S = I$, $d\phi_I(E) = 0$ by definition of E and so we would expect that $\nabla\phi(I) = 0$. Now, in order to find $\nabla\phi$, we use the relation $d\phi_S(SE) = g(\nabla\phi(S), SE)_S = g(S^{-1}\nabla\phi(S), E)_I = -\frac{1}{2}\mathrm{Tr}(S^{-1}\nabla\phi(S)E)$. However, if we were to naïvely write $\frac{1}{2}S^{-1}\nabla\phi(S) = cS$, it implies that $\nabla\phi(S) = 2cS^2$. This is clearly not correct since S^2 is not necessarily in the Lie algebra and moreover, $\nabla\phi(I) \neq 0$. To ensure that $S^{-1}\nabla\phi(S)$ lies in the Lie algebra, we add in the skew symmetric constraint $(S^{-1}\nabla\phi(S))^\intercal = -S^{-1}\nabla\phi(S)$.

The solution to this constrained problem is then

$$\nabla\phi(S) = cS(S - S^{\intercal}),$$

and it is clear that $S^{-1}\nabla\phi(S) \in \mathfrak{so}(n)$.

One may verify this claim by substituting it back into the inner product

$$\begin{aligned} d\phi_S(SE) &= g(\nabla\phi(S), SE)_S, \\ &= g(S^{-1}\nabla\phi(s), E)_I, \\ &= \frac{c}{2}\mathrm{Tr}((S^{\intercal} - S)E), \\ &= -c\mathrm{Tr}(SE). \end{aligned}$$

If one were interested in the case where S_0 is not the identity, then this result can be re-obtained by applying the left action of S_0 onto S in $S^{-1}\nabla\phi(S)$, yielding

$$\nabla\phi(S) = cS(S_0S - (S_0S)^{\intercal}). \tag{4.22}$$

Proposition 4.3.1. *For $\phi = -c\mathrm{Tr}(S_0S)$, we have the following lower bound on the Hessian:*

$$\mathrm{Hess}^{\phi} \geq -2cg.$$

Proof. Without loss of generality, we set $S_0 = I$.

To first calculate the Hessian, we define $\gamma(t) = Se^{tE}$ so that $\gamma(0) = S$ and $\dot{\gamma}(0) = SE$. The Hessian is then defined as

$$\begin{aligned} \mathrm{Hess}^{\phi}(SE, SE) &= \frac{d^2}{dt^2}\phi(\gamma(t))\bigg|_{t=0} \\ &= \frac{d}{dt} - c\mathrm{Tr}(SEe^{tE})\bigg|_{t=0} \\ &= -c\mathrm{Tr}(SE^2e^{tE})\bigg|_{t=0} \\ &= -c\mathrm{Tr}(SE^2). \end{aligned}$$

For $i, j = 1, ..., n$, write $\tilde{E}_{ij}$ for the $n \times n$ matrix whose (i, j)th entry is 1, zero otherwise. Then, for $i < j$, $\tilde{E}_{ij}\tilde{E}_{ij} = 0$ and $\tilde{E}_{ij}\tilde{E}_{ji} = \tilde{E}_{ii}$. We define the skew-symmetric matrix E with entries $E_{ij} = \tilde{E}_{ij} - \tilde{E}_{ji}$ for $i, j = 1, ..., n$. Because our analysis on the Hessian strictly involves the trace, we shall set our focus on the diagonal entries of SE^2. If S_{ij} is the (i, j)th entry of S, $(SE^2)_{kk} = S_{ki}E_{ij}E_{jk}$. Expanding E^2 in terms of $\tilde{E}$,

$$\begin{aligned} E_{ij}E_{jk} &= (\tilde{E}_{ij} - \tilde{E}_{ji})(\tilde{E}_{jk} - \tilde{E}_{kj}) \\ &= \tilde{E}_{ij}\tilde{E}_{jk} - \tilde{E}_{ij}\tilde{E}_{kj} - \tilde{E}_{ji}\tilde{E}_{jk} + \tilde{E}_{ji}\tilde{E}_{kj}. \end{aligned}$$

However, by definition of $\tilde{E}$,

$$\tilde{E}_{ij}\tilde{E}_{kl} = \begin{cases} \tilde{E}_{il} & j = k, \\ 0 & \text{otherwise.} \end{cases}$$

and therefore, $E_{ij}E_{jk} = \tilde{E}_{ik}$. The Hessian can now be reformulated as

$$\mathrm{Hess}^{\phi}(SE, SE) = -c\mathrm{Tr}(SE^2) = -cS_{ki}\tilde{E}_{ik} \geq c\tilde{E}_{ik} = c\mathrm{Tr}(E^2)$$

since $|S_{ij}| \leq 1$ by the restriction that $S^{\intercal}S = I$ for any rotation matrix. In terms of the inner product,

$$\mathrm{Hess}^{\phi}(SE, SE) \geq c\mathrm{Tr}(E^2) = -2cg(E, E)_I. \tag{4.23}$$

□

Note that this result is invariant to the choice of S_0, since SO(n) is closed under left multiplication.

We are now able to find what values of c satisfy the Bakry-Émery-Ricci criterion. Combining Proposition 4.3.1 together with (4.23) we obtain

$$\mathrm{Ric}(SE, SE) + \mathrm{Hess}^{\phi}(SE, SE) \geq \left(\frac{n-2}{2} - 2c\right) g(SE, SE)_S$$

and so the right hand side is positive if and only if $\frac{n-2}{2} > 2c$, i.e. $c < \frac{n-2}{4}$. In the previous chapter, we discussed and constructed a Stein method for the von-Mises distribution, although this did not involve any sort of diffusions nor infinitesimal generators. In the case when $n = 2$, $\mathrm{SO}(2) \cong \mathbb{S}^1$ and $p_{MVM}(S) \propto e^{2c\cos(\theta)}$ which is exactly the von-Mises density. Furthermore, the failure to satisfy the Bakry-Émery-Ricci criterion coincides with the violation of the criterion when $M = \mathbb{S}^1$ and $\phi = -k\cos(x - \mu)$. This is because $\mathrm{Ric}(E, E) = 0$ and so we would need a negative value for c which violates the requirement on $c > 0$. Moreover, the computation for the Hessian is also consistent since we parametrised $k = 2c$ for the von-Mises distribution.

The first comparison that shall be examined will be between two matrix von-Mises Fisher distributions with equal location parameters, but different scale parameters.

Proposition 4.3.2. *Let* $X \sim \mathrm{MVM}(S_0, c_1)$ *and* $Y \sim \mathrm{MVM}(S_0, c_2)$ *be matrix von-Mises Fisher distributions on* $\mathrm{SO}(n)$*. Assuming that* $0 < c_1, c_2 < \frac{n-2}{4}$*,*

$$d_W(Y, X) \leq \frac{|c_2 - c_1|}{2\kappa} \mathbb{E}_X[\sqrt{n - \mathrm{Tr}((S_0X)^2)}].$$

Proof. With an abuse of notation, define the random variables $X \sim p_1$ and $Y \sim p_2$ where $p_1(S) \propto \exp(c_1 \mathrm{Tr}(S_0 S))$ and $p_2(S) \propto \exp(c_2 \mathrm{Tr}(S_0 S))$. Further define $\phi := -\log p_1 = -c_1 \mathrm{Tr}(S_0 S) + C_1$ and $\psi := -\log p_2 = -c_2 \mathrm{Tr}(S_0 S) + C_2$ where C_1 and C_2 are constants — we need not worry about what the constants are since we are going to be applying the gradient operator to ϕ and ψ. Then by applying the result in Equation (4.22) we have that $\nabla\phi(S) = -c_1 S(S_0 S - (S_0 S)^{\intercal})$ and

$\nabla\psi(S) = -c_2 S(S_0 S - (S_0 S)^\intercal)$. We next calculate the norm of $\nabla\phi - \nabla\psi$;

$$\begin{aligned}|\nabla\phi(S) - \nabla\psi(S)|^2 &= \frac{(c_2 - c_1)^2}{2}\mathrm{Tr}\big((S_0 S - (S_0 S)^\intercal)^\intercal(S_0 S - (S_0 S)^\intercal)\big)\\ &= \frac{(c_2 - c_1)^2}{2}\mathrm{Tr}\big(((S_0 S)^\intercal - S_0 S)(S_0 S - (S_0 S)^\intercal)\big)\\ &= \frac{(c_2 - c_1)^2}{2}\mathrm{Tr}(2I - (S_0 S)^2 - ((S_0 S)^\intercal)^2)\\ &= \frac{(c_2 - c_1)^2}{2}\big(2n - 2\mathrm{Tr}((S_0 S)^2)\big).\end{aligned}$$

The expectation of the square root of this quantity is simply

$$\mathbb{E}_X[|\nabla\phi(S) - \nabla\psi(S)|] = \frac{|c_2 - c_1|}{\sqrt{2}}\mathbb{E}_X[\sqrt{2n - 2\mathrm{Tr}((S_0 X)^2)}].$$

Therefore, using Theorem 4.0.1, an upper bound on the Wasserstein metric is

$$d_W(Y, X) \leq \frac{|c_2 - c_1|}{2\kappa}\mathbb{E}_X[\sqrt{n - \mathrm{Tr}((S_0 X)^2)}]$$

for $c_1, c_2 < \frac{n-2}{4}$. □

If $c_1 = c_2$ then clearly $d_W(Y, X) = 0$. Since all of the eigenvalues of special orthogonal matrices lie on the unit circle in $\mathbb{C}$, all real parts lie between -1 and 1 and therefore, the absolute value of the trace will be no greater than n. Hence, a less sharp bound is

$$d_W(Y, X) \leq \frac{|c_2 - c_1|}{2\kappa}\sqrt{2n}.$$

One can also use the above result to generate an upper bound between a matrix von-Mises Fisher distributions and the Haar (uniform) measure on SO(n). Since the Haar measure Z on SO(n) is such that $S_0 Z$ is also distributed as a Haar measure for any $S_0 \in SO(n)$, the subsequent corollary follows from Proposition 4.3.2 when one of c_1, c_2 is 0.

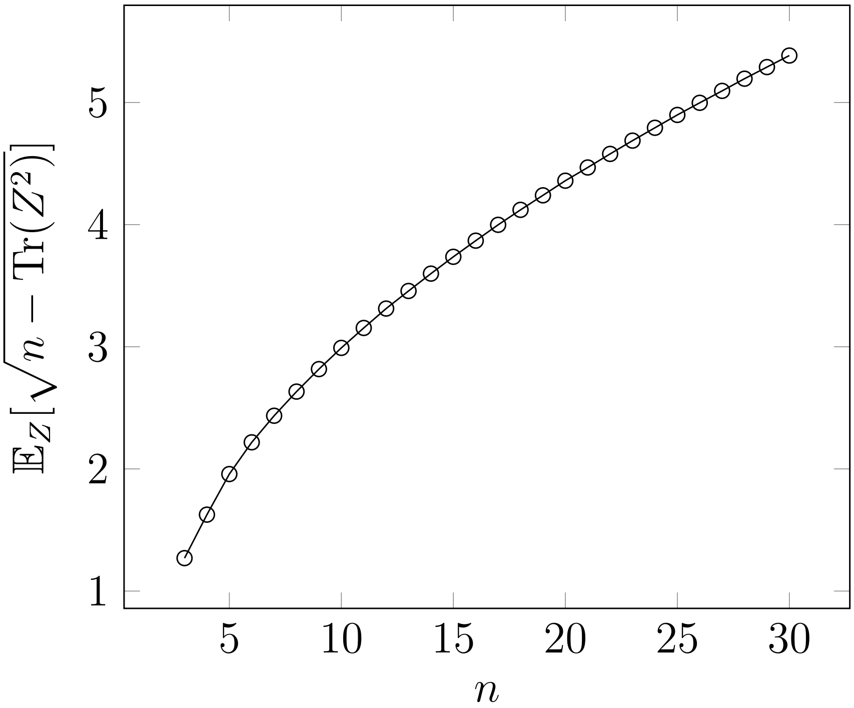

Figure 4.2: A plot of the expected value in Equation (4.24) against the dimension n. The circles are the data points for the value of the expectation and the line is an interpolation between each point.

Corollary 4.3.3. *Let X follow a matrix von-Mises distribution on* $\mathrm{SO}(n)$ *and Z the Haar measure on* $\mathrm{SO}(n)$*. Then we have the following bound on the Wasserstein distance:*

$$d_W(Y, Z) \leq \frac{c}{2\kappa}\mathbb{E}_Z[\sqrt{n - \mathrm{Tr}(Z^2)}]. \tag{4.24}$$

Figure 4.2 displays simulation results for computing $\mathbb{E}_Z[\sqrt{n - \mathrm{Tr}(Z^2)}]$ in Equation (4.24). A basic Monte Carlo integration method was employed with 100,000 variates of the Haar measure on $\mathrm{SO}(n)$ drawn for each n.

4.4 Comparison on $\mathcal{P}_n$

We now end the chapter with a brief construction and analysis of $\mathcal{P}_n$, the space of symmetric positive definite $n \times n$ matrices. $\mathcal{P}_n$ is a homogeneous space of $GL(n)$; identified with the quotient $GL(n)/O(n)$. The tangent space at a point $P \in \mathcal{P}_n$ is identified as $\{P\} \times \mathrm{Sym}(n) =: T_P\mathcal{P}_n$ where $\mathrm{Sym}(n)$ is the space of symmetric

$n \times n$ matrices. For brevity, we shall just associate $T_P\mathcal{P}_n$ with Sym(n) alone.

The manifold $\mathcal{P}_n$ is of particular interest to statistics. Covariance matrices of multivariate distributions are symmetric positive definite matrices and so models that involve the comparison of covariance matrices could use this framework.

We shall attach the affine-invariant metric to $\mathcal{P}_n$, which will give it a Riemannian structure. For $A, B \in T_P\mathcal{P}_n$, the affine-invariant metric is

$$g(A, B) = \frac{1}{2}\mathrm{Tr}(P^{-1}AP^{-1}B).$$

With this particular choice of metric, the manifold $\mathcal{P}_n$ has negative sectional curvature. The scalar curvature has also been explicitly computed in [And03] as

$$S = -\frac{1}{8}n(n-1)(n+2).$$

The sectional curvature, however, is not constant. This makes computing a lower bound for the Ricci curvature more challenging. Explicitly computed quantities for the Christoffel symbols and metric tensor in dimension 3 can be found in [MZ11].

The affine-invariant metric, as opposed to the Euclidean metric, gives $\mathcal{P}_n$ negative curvature. The most important distinction, however, is that the Euclidean metric makes $\mathcal{P}_n$ a closed manifold, i.e. with a boundary. Therefore we may not apply the framework of [LLBF22] in this case. Moreover, the exponential map and geodesic paths are readily available with the affine-invariant metric.

The Ricci curvature tensor has been previously found in [DP18, Proposition 2.4] to satisfy

$$\mathrm{Ric}_P(X, Z) = \frac{1}{4}\mathrm{Tr}(P^{-1}X)\mathrm{Tr}(P^{-1}Z) - \frac{n}{4}g_{\mathcal{P}_n}(X, Z),$$

for $P \in \mathcal{P}_n$ and $X, Z \in T_P\mathcal{P}_n$. For our purposes, it is sufficient to perform analysis in the case that $X = Z$ — the proof requiring the Bakry-Èmery-Ricci criterion only needs the two arguments X, Z to be equal. Thus:

Lemma 4.4.1. *Let $g_{\mathcal{P}_n}$ and $\mathrm{Ric}_{\mathcal{P}_n}$ denote the metric and Ricci curvature tensors of $\mathcal{P}_n$ respectively, then*

$$\mathrm{Ric}_{\mathcal{P}_n}(X,X) \geq -\frac{n}{4} g_{\mathcal{P}_n}(X,X)$$

for $X \in TM$.

We now focus our attention to the Riemannian-Gaussian distribution on $\mathcal{P}_n$. This distribution was introduced for $\mathbb{H}^3$ in Equation (4.13). In contrast with the work in Section 4.2, we shall not regard $\mathcal{P}_n$ as a spherically symmetric manifold, and so we will rewrite the density (4.13) in a more explicit form;

$$p_{\mathrm{RG}}(P) = \frac{1}{C}\exp\big(-c\rho^2(P,P_0)\big), \quad c > 0, P_0 \in \mathcal{P}_n \tag{4.25}$$

where C is the normalizing constant $C = \int_{\mathcal{P}_n} e^{-c\rho^2(P,P_0)} d\mathrm{vol}$

We now need to find two things in order to apply Theorem 4.0.1: a form for $\nabla\rho^2(P,P_0)$ and a lower bound on Hess^{ρ^2}.

We first begin with $\nabla\rho^2$. Define the path $\gamma : [0,1] \to \mathcal{P}_n$ as the geodesic connecting two points $\gamma(0) = P$ and $\gamma(1) = P_0$. The geodesic distance ρ connecting P and P_0 is then minimized in the direction of $\dot\gamma(0)$ — the negative curvature ensures this is unique. Because the gradient operator ∇ points in the direction of maximization, it is the case that $\nabla\rho(P,P_0) = -\dot\gamma(0)$. Moreover, it is also the case that $\dot\gamma(0) = \exp_P^{-1}(P_0)$ since there is no cut locus. Therefore, we have the following equation for the gradient

$$\nabla\rho^2(P,P_0) = -2\exp_P^{-1}(P_0). \tag{4.26}$$

One can go further on to explicitly calculate the geodesic path as

$$P(t) = P^{1/2}\exp\big(t\mathrm{Log}(P^{-1/2}P_0P^{-1/2})\big)P^{-1/2}$$

and hence

$$\exp_P^{-1}(P_0) = P^{1/2}\mathrm{Log}(P^{-1/2}P_0P^{-1/2})P^{1/2},$$

where Log is the matrix logarithm — see [MZ11].

We now move on to finding a bound for the Hessian. Since $\mathcal{P}_n$ has negative sectional curvature, the Cartan-Hadamard theorem states that the exponential map is everywhere injective. Consequently, the distance function is a convex function and hence $\mathrm{Hess}^{\rho^2} \geq 0$. Using this fact, we pursue use the Hessian comparison theorem to find a lower bound on the Hessian. Our manifold of comparison shall be the Hyperbolic space $\mathbb{H}^{n(n+1)/2}(-1)$. We endow $\mathbb{H}^{n(n+1)/2}(-1)$ with the canonical metric as described in Section 4.2. We scale the metric $g_{\mathcal{P}_n}$ in such a way that the sectional curvatures $-K_{\mathcal{P}}$ of $\mathcal{P}_n$ satisfy $-K_{\mathcal{P}} < -1$.

We are now at a point where an application of the Hessian comparison theorem is possible.

Theorem 4.4.2 (Hessian Comparison Theorem)**.** *Let (M, g), (N, h) be Riemannian manifolds. Let $\gamma : [0, 1] \to M$ and $\tilde{\gamma} : [0, 1] \to N$ denote geodesics on M and N respectively. Denote by $K(t)$ and $\tilde{K}(t)$ the sectional curvature of M and N restricted to the geodesic at time t. Suppose that $\tilde{K}(t) \leq K(t)$ holds for all $t \in [0, 1]$. Suppose further that for vector fields $X_p \in T_pM$ and $Y_q \in T_qM$ for $p = \gamma(a)$ and $q = \tilde{\gamma}(a)$ $a \in [0, 1]$,*

$$g(X_p, \dot{\gamma}(a)) = g(Y_q, \dot{\tilde{\gamma}}(a)),$$

and $|X_p| = |Y_q|$. Then

$$\mathrm{Hess}^{\rho^2}(X_p, X_p) \leq \mathrm{Hess}^{\tilde{\rho}^2}(Y_q, Y_q).$$

By the Hessian comparison theorem, for vectors $X \in T_P\mathcal{P}_n$, $Y \in T_y\mathbb{H}^{n(n+1)/2}$,

$$\mathrm{Hess}_{\mathbb{H}}^{\rho^2}(Y,Y) \leq \mathrm{Hess}_{\mathcal{P}}^{\rho^2}(X,X)$$

such that $g_{\mathcal{P}_n}(X,X) = g_{\mathbb{H}}(Y,Y)$. Therefore, on applying the Hessian comparison theorem with the lower bound on $\mathrm{Hess}_{\mathbb{H}}^{\rho^2}$, we obtain

$$\begin{aligned}\mathrm{Hess}_{\mathcal{P}_n}^{\rho^2}(X,X) &\geq \mathrm{Hess}_{\mathbb{H}}^{\rho^2}(Y,Y),\\ &\geq 2g_{\mathbb{H}}(Y,Y) = 2g_{\mathcal{P}_n}(X,X)\end{aligned}$$

ultimately giving

$$\mathrm{Hess}^{c\rho^2} \geq 2cg. \tag{4.27}$$

On combining Lemma 4.4.1 and Equation 4.27 we have the following result:

Lemma 4.4.3. *For $\phi = c\rho^2(P,P_0)$, the sufficient condition is satisfied for values of c such that $c > \frac{n}{8}$.*

Remark. Since the Ricci curvature is bounded from below and the Hessian is positive semi-definite, the diffusion process

$$\begin{aligned}dX_t &= \Xi_t(X_t)\circ dB_t - \frac{1}{2}\nabla\rho^2(X_t)dt, \quad X_0 = x \in \mathcal{P}_n, \qquad (4.28)\\ d\Xi_t &= H_\Xi dX_t\end{aligned}$$

never exits $\mathcal{P}_n$ — the Bakry-Émery-Ricci tensor is bounded from below. Therefore, in theory, one could use the SDE (4.28) with a suitable numerical scheme to sample from the density (4.25) by allowing the diffusion to go on long enough. Sampling from the density (4.25) has already been explored in [SBBM17] in which they employ a rejection sampling algorithm to do so.

With Lemma 4.4.3, we can utilise Theorem 4.0.1 to bound the Wasserstein metric above for two different Riemannian Gaussian distributions. Let $X \sim p_\phi$ and $Y \sim p_\psi$ where $\phi(P) = c_1\rho^2(P,P_1)$ and $\psi(P) = c_2\rho^2(P,P_2)$ are the exponents in the pdf (4.25). Using Equation (4.26)

$$\nabla(\phi - \psi)(P) = -2(c_1 \mathrm{exp}_P^{-1}(P_1) - c_2 \mathrm{exp}_P^{-1}(P_2))$$

and so

$$\begin{aligned} d_W(X,Y) &\leq \frac{1}{2\kappa}\mathbb{E}_X\left[\left|\sqrt{\mathrm{Tr}(X^{-1}(c_1\mathrm{exp}_X^{-1}(P_1) - c_2\mathrm{exp}_X^{-1}(P_2)X^{-1}(\mathrm{exp}_X^{-1}(P_1) - \mathrm{exp}_X^{-1}(P_2))}\right|\right] \\ &= \frac{1}{2\kappa}\mathbb{E}_X\left[\left|\sqrt{\mathrm{Tr}\big((X^{-1/2}(c_1\mathrm{exp}_X^{-1}(P_1) - c_2\mathrm{exp}_X^{-1}(P_2))X^{-1/2})^2\big)}\right|\right] \end{aligned}$$

for some $\kappa > 0$. In the special case when $P_1 = P_2$, we may further simplify the right hand side:

$$\nabla(\phi - \psi) = (c_1 - c_2)\nabla\rho^2(P, P_1) = 2(c_1 - c_2)\dot{\gamma}(0)$$

where γ is the intervening geodesic connecting P and P_1. Therefore we have the following simplification for the upper bound of the Wasserstein metric,

$$\begin{aligned} d_W(X,Y) &\leq \frac{|c_1 - c_2|}{2\kappa}\mathbb{E}_X[|2\dot{\gamma}(0)|], \\ &= \frac{|c_1 - c_2|}{\kappa}\mathbb{E}_X[\rho(X, P_1)]. \end{aligned}$$

4.5 Conclusion

To conclude, we briefly summarise the key techniques and findings of this chapter.

For calculating bounds on the Wasserstein metric by following the method presented in [LLBF22], the main hurdle that one encounters is showing that the Bakry-Émery-Ricci criterion is satisfied. The majority of the work in this chapter was showing this for each of the distributions we wished to compare. Particularly, calculating bounds on the Hessian was the most challenging aspect. We used a selection of methods overcome this: using properties of spherically symmetric manifolds and exact expressions of the Hessian of the distance function; calculating the Hessian from its definition in terms of geodesics, as is in Equation (2.1); applying the Hessian comparison theorem. Together with 4.0.1, we were able to generate completely new and unseen bounds between a number of probability measures on a variety of spaces.

It is also possible to show the Bakry-Émery-Ricci criterion is satisfied directly by using local coordinates and brute-force calculating all necessary quantities like the Christoffel symbols and Ricci curvature, however this is a very tedious and time consuming task that is a lot less elegant. Loss of generality for dimension will also occur, since one will have to select a coordinate chart on a fixed dimensional manifold to work in. Although it could be reasonable to do in a low dimensional manifold setting, for example $\mathbb{H}^2$ where the number of Christoffel symbols is low and a global coordinate chart exists.

Like with previous research in distributional convergence, diffusion approach [LLBF22] can be used to assist in the generation of random variables whose variates are typically hard to draw from. For example, if we wanted to draw variates from a matrix von-Mises distribution with small scaling parameter c, we could instead draw a Haar variate and conclude that this is a good approximation by Equation (4.24).

One direction that could lead to fruitful application would be to modify the framework of [LLBF22] so that it is possible to compare a continuous distribution to a discrete distribution on the manifold, the key motivation being that an empirical distribution of data on a manifold is discrete. One idea is to generate a Markov chain whose invariant distribution is the empirical distribution and to find the Stein operator that way. However, care is needed when discretising the manifold itself.

In the next chapter, we will be extending the underlying framework of the method used in this chapter to the case where a boundary is present.

Chapter 5

Stein's Method on Manifolds with Boundary

The Stein's method presented in Le et. al. [LLBF22] provided a strong foundation on which one can construct a Stein's method for a general Riemannian manifold. However, for practical data applications, one may decide to only use a closed subset of a manifold as the sample space instead of the full space. A brief example would be taking the positional data of thrown darts at a dart board. The entirety of $\mathbb{R}^2$ isn't needed and so one may instead model the data on the closed ball $\{x \in \mathbb{R}^2 : |x|^2 \leq 1\}$, with the data having both radial and angular components. The underlying assumption that the manifold does not contain a boundary is then violated and thus, the method is inapplicable. To overcome such a problem, we introduce a local time process into the unreflected Feller diffusion on an open manifold. This allows one to construct a process which is Feller in the interior of a manifold with boundary and does not leave said manifold. This inclusion of a local time term, however, also presents a multitude of additional problems one must work through if one wants to pursue a construction similar to that in [LLBF22].

Our strategy for this chapter is to extend the theory from the boundary-less case in a careful manner so that we may adapt the results from this case to the boundary case. We decompose this chapter into 6 sections with 3 subsections. We first begin with setting up the necessary definitions and assumptions required for the chapter. In Section 5.2, we then go through the process of identifying the correct Stein operator for our target measure. Following this, we introduce the coupling theory presented in [LLBF22] and extends it with the inclusion of local time terms in Section 5.3. After that, in Section 5.4, we construct the Stein equation, and prove the particular form of the solution as well as bound the solution and its first derivative. The next section, we break up into three more digestible subsections with the overall aim of Section 5.5 to bound the second derivative of the solution to the Stein equation. We briefly introduce the Weitzenböck formula in Subsection 5.5.1 as a necessary tool to prove a Bismut-Elworthy-Li formula in the subsequent subsection. We use Subsection 5.5.2 as a means to arrive at this result by using Damped Stochastic Parallel Displacement. This is a flow similar to that of the derivative flow that leads us to the desired result. In Section 5.5.3 we use this Bismit-Elworthy-Li formula and Damped Stochastic Parallel Translation to bound the second derivative. To conclude the chapter, Section 5.6 goes through bounding the Wasserstein metric, and an example is given with the von-Mises Fisher distribution on the small cap of $\mathbb{S}^n$.

5.1 Preliminary Foundations

Let M be an n-dimensional Riemannian manifold with C^∞ boundary with metric g and we shall equip (M, g) with the Levi-Civita connection D. Denote by $\overset{\circ}{M}$ the interior and by ∂M the boundary of M. We use $\Pi_{x,v}$ to represent the parallel transport of a vector field over a geodesic from a point x propagated in the direction of v. If a value $y \in M$ is within the radius of injectivity of $x \in M$, i.e. $\exp_x(tv)$ is a geodesic for $t \in [0, 1]$, and $y = \exp_x(v)$ then we denote this mapping as $\Pi_{x,v} : T_xM \to T_yM$. We assume that M and its boundary ∂M are connected.

We review some important definitions from [Kro79] on convex manifolds:

Definition 5.1.1. Given a Riemannian manifold N and a set $C \subset N$, C is said to be *convex* if, for any point $p \in \bar{C}$, there is a number $e(p)$ with $0 < e(p) < r(p)$ such that $C \cap B_{e(p)}(p)$ has the property that between any two points in C there is a unique minimal geodesic in N completely contained in $C \cap B_{e(p)}(p)$ which joins these points. Here, $r(p)$ is known as the radius of convexity of N at p and we denote $B_r(p)$ as the geodesic ball of radius r centred at p in N. This is taken to be the smallest number such that $C \cap B_{r(p)}(p)$ contains non-unique geodesics.

Example 5.1.2. Suppose $N = \mathbb{S}^n$, then the spherical cap $M_c = \{(\rho, \theta) \in N : \rho \in [0, c]\}$ ($\mathbb{S}^n$ is coordinatised as $\rho \in [0, \pi]$, $\theta \in \mathbb{S}^{n-1}$), where $c \in [0, \pi/2]$, is a submanifold with boundary of N that has radius of convexity $\pi/2$ — we inherit the canonical metric of N. If we restrict c to be less than $\pi/2$ then the submanifold with boundary has no cut points and every geodesic in M_c is unique.

This definition shows that C can be regarded as an embedded topological manifold of N with smooth, totally geodesic interior $\mathring{C}$ and (possibly non-smooth) boundary ∂C. If ∂C is indeed smooth, then C is a smooth submanifold with boundary of N. The boundary of such a set is called a convex curve.

It is clear that any geodesic ball with small enough radius will be geodesically convex. For example, taking $r < \frac{1}{2}\text{inj}(N)$, $B_p(r)$ for $p \in N$ contains unique geodesics only and with boundary equal to $\{x \in N : \rho(x, p) = r\}$.

Due to the fact that we have embedded M within N, it is a consequence that we shall be inheriting the metric structure and geometry of N for M. Going forward, we shall assume that there exists a manifold N such that M is a convex submanifold with boundary of M. A useful fact for identifying such manifolds is that a compact Riemannian manifold with boundary ∂M is convex if and only if the second fundamental form $\mathbb{II}$ of the outward normal vector field in the normal bundle along ∂M is positive semidefinite [Kro79].

As before in the general manifold case [LLBF22], our interests lie in being able to construct a Stein method for a probability measure of the form

$$d\mu_\phi = \frac{1}{C_\phi} e^{-\phi} d\mathrm{vol}, \tag{5.1}$$

where $C_\phi = \int_M e^{-\phi} d\mathrm{vol} < \infty$ and the measure has support over the entire space M. We assume that $\phi \in C^3(M)$ and that $\nabla\phi$ is Lipschitz. For the remainder, we shall write that the random variable $X \sim \mu_\phi$ if X has probability measure given by (5.1). We note that the measure of the boundary is 0, $\mu_\phi(\partial M) = 0$, since ∂M is codimension 1 with respect to M.

We define the Feller diffusion process $\{X_t\}_{t\in\mathbb{R}^+}$ as the solution to the reflected stochastic differential equation

$$dX_t = dB_t^M - \frac{1}{2}\nabla\phi(X_t) + \nu(X_t)dL_t, \quad X_0 = x \in M. \tag{5.2}$$

Here, the process $\{B_t^M\}_{t\in\mathbb{R}^+}$ denotes an M-valued Brownian motion, $\nu(x)$ is the inward pointing normal vector field at x, and $\{L_t\}_{t\in\mathbb{R}^+}$ is the so called local time of the process $\{X_t\}_{t\in\mathbb{R}^+}$. See Figure 5.1 for a pictorial representation of the diffusion. This is a continuous, non-decreasing process that satisfies

$$\int_0^t \mathbb{I}_{\mathring{M}}(X_s)dL_s = 0.$$

In other words, L_t increases only on the boundary. Consequently, the process X_t behaves as an overdamped Langevin diffusion in the interior. Moreover, $\{L_t\}_{t\in\mathbb{R}^+}$ is a bounded variation process and therefore $dL_t^2 = dB_t^M dL_t = dL_t dt = 0$

Remark. It is possible to extend Paul Lévy's notion of "mesure du voisinage" (cf. [GH80] or [Xia98, pp. 386]) from general Borel sets of $\mathbb{R}^d$ to manifold with boundary by considering an ϵ-tubular neighbourhood around the boundary. By shrinking the diameter ϵ of the tubular neighbourhood, a quotient between the occupation time of the process and the induced volume measure of the boundary can be produced, yielding a local time process.

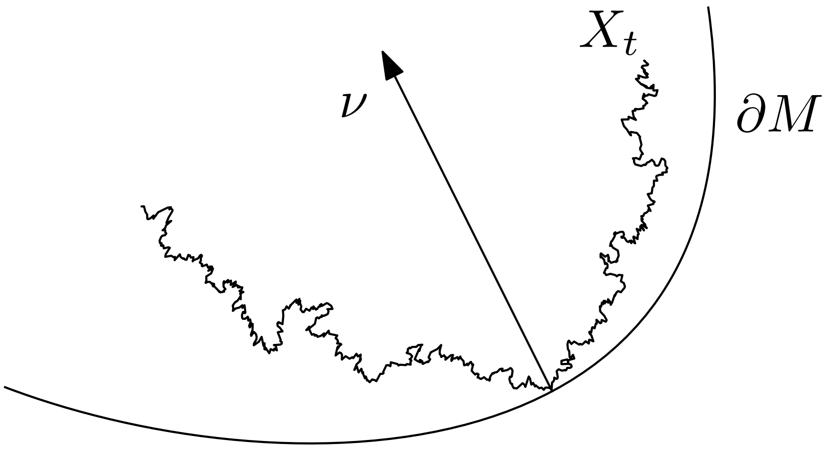

Figure 5.1: Diagram of a reflected diffusion X_t in M, reflecting about the normal vector ν.

This particular type of SDE has been previously used to describe reflected Brownian motions with drift [Wan94]. Note that this diffusion does not leave M a.s. because the local time term $\nu(X_t)dL_t$ reflects the process inward once it hits the boundary.

The inclusion of the local time term is a key step for our extension from manifold to manifold with boundary. In the interior $\overset{\circ}{M}$, the diffusion (5.2) is Feller and has the usual infinitesimal generator which we require to be the Stein operator.

We induce a uniform measure on the ∂M from M by the following procedure: Let $F_1 = (ue_1, ue_2, ..., ue_{n-1}, \nu)$ be an oriented orthonormal frame in T_xN where ν is the inward normal vector of M — $\nu \in N_x(M)$. Then since $M \subset N$, $F_2 = (ue_1, ue_2, ..., ue_{n-1})$ is an oriented orthonormal frame in T_xM. Since F_1 is orthonormal, it is the case that $d\text{vol}(F_1) = 1$ (cf. [Lee18, pp. 432]). We now see that by the interior product

$$\iota_\nu d\text{vol}(F_2) = d\text{vol}(F_1) = 1$$

is also a volume form, but instead, over the boundary ∂M. We then define $d\text{vol}(\partial M)(x) = d\text{vol}(\nu) = \iota_\nu d\text{vol}$ for the unit normal vector $\nu \in N_x(M)$, the normal tangent space. The interior product ι_ν acting on a p-form ω maps to the space of $(p-1)$-forms.

Since M is embedded within N, if we assume for the time being that $\phi \in C^3(N)$ (instead of just $C^3(M)$), we can construct an unreflected process which exists on N by eliminating the reflection term in (5.2). The infinitesimal generator for such a diffusion takes the form

$$\mathcal{A} = \frac{1}{2}\Delta_N - \frac{1}{2}g(\nabla\phi, \nabla) \tag{5.3}$$

which acts on a suitable family of functions so that $|\mathbb{E}[\mathcal{A}f]|$ is finite. This operator then generates the Itô diffusion

$$dU_t = dB_t^N - \frac{1}{2}\nabla\phi(U_t)dt, \quad U_0 = u \in N. \tag{5.4}$$

Remark. The reason we require the reflection component in the SDE (5.2) is because we eventually want to show that the stationary/invariant distribution of $\{U_t\}_{t\in\mathbb{R}^+}$ is the measure μ_ϕ. Since the measure is only defined on M, we must restrict the SDE to M to avoid problems of convergence in its ergodic limit. For example, suppose we were interested in finding a process whose invariant distribution is exponential with rate λ on $\mathbb{R}$. A potential candidate that we might choose is a Brownian motion with drift, $dU_t = dB_t - \lambda dt$ because the infinitesimal generator of such a process generates mean zero functions under the exponential distribution. The problem, however, lies in the fact that a Brownian motion with drift and diverges to $-\infty$ a.s. Therefore, the stationary distribution doesn't exist and so we must restrict the process to $\mathbb{R}^+$ via a reflection at 0.

We shall assume that the unreflected process (5.4) is stochastically complete on N. By this we mean the process does not exit N in finite time. An equivalent condition for stochastic completeness to hold is to show that [Bak86], for some $\kappa > 0$,

$$\text{Ric} + \text{Hess}^\phi \geq -\kappa g. \tag{5.5}$$

Now, because U_t does not leave N, it is clear that X_t does not leave M by virtue of the reflection term in (5.2).

In order to construct a Stein method on M, we must demand a condition stronger than (5.5) to properly develop the framework. This new assumption is not too disimilar to the inequality (5.5), but instead of requiring the Bakry-Émery-Ricci tensor to be bounded from below, we demand that it is positive-definite:

$$\mathrm{Ric} + \mathrm{Hess}^{\phi} \geq 2\kappa g \tag{5.6}$$

for some $\kappa > 0$.

If we take M to be a convex subset of $\mathbb{R}^n$ with flat metric, then the sufficient condition reduces to $\mathrm{Hess}^{\phi} \geq 2\kappa g$. Or in other words, for a vector $u \in \{v \in \mathbb{R}^n : |v| = 1\}$, $u^{\intercal}\mathrm{Hess}^{\phi}u \geq 2\kappa$. This condition, unsurprisingly, is equivalent to the requirement on ϕ in [MG16]. In the terminology of Mackey and Gorham [MG16], ϕ is strongly 2κ-concave.

The Brownian motion B_t^M of X_t in (5.2) is constructed via horizontal lift. That is to say, for an n-dimensional Brownian motion $\{B_t\}_{t\in\mathbb{R}^+}$

$$\begin{aligned} dX_t &= \Xi_t(X_t) \circ dB_t - \frac{1}{2}\nabla\phi(X_t)dt + \nu(X_t)dL_t, \quad X_0 = x \in M, \\ d\Xi_t &= H_{\Xi} \circ dX_t, \quad \Xi_0 = \xi \in \mathcal{O}(M), \end{aligned} \tag{5.7}$$

where Ξ is a lift of X_t to the orthonormal frame bundle $\mathcal{O}(M)$ and H_{Ξ} is the horizontal lift. We have also rewritten $\Xi_t = \Xi_t(X_t)$ for brevity. It is known that (cf. [Wan14, Chapter 3]) if the drift function is C^1, then the process (5.7) has a unique solution up to explosion time. Since ϕ is C^3 and our manifold with boundary is stochastically complete, we need not ponder about such problems.

For later use, we define the following Lipschitz constants: For a function $g \in C^k(M)$,

$$C_i(g) = \frac{\left\|D^i g(x) - \Pi_{\gamma_{x,y}}(D^i g(y))\right\|_{\mathrm{op}}}{\rho(x,y)}, \quad i = 0, .., k, \tag{5.8}$$

where D^i is the covariant derivative applied i times. In this, $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm, $\gamma_{x,y}$ denotes any possible geodesic from y to x and therefore, $\Pi_{\gamma_{x,y}}$ is notation for the parallel transport of a vector from T_yM to T_xM along $\gamma_{x,y}$. We also define $D^0 := \mathrm{Id}$ and we note that the operator norm of a function is the infinity norm $\|\cdot\|_\infty$. We will find that the two quantities $C_0(g) = \|dg\|_{\mathrm{op}}$ and $C_1(g) = \|\mathrm{Hess}^g\|_{\mathrm{op}}$ will be of great importance to us later on in the chapter, specifically in relation to bounding the solution to the Stein equation.

5.2 The Stein Operator

As with all constructions of Stein's method, the first step in the general procedure will be to find an operator $\mathcal{L}$ on M such that $\mathbb{E}_{\mu_\phi}[\mathcal{L}f] = 0$ for some space of functions f. For the diffusion approach, $\mathcal{L}$ is the infinitesimal generator of the stochastic process whose invariant distribution is μ_ϕ. Compared to ordinary diffusions, it is not as easy to find the infinitesimal generator, and then further show that μ_ϕ is the invariant distributions.

In the non-reflected case, an alternative way to formulate the problem is to instead find an operator $\mathcal{A}$ such that

$$f(X_t) - \int_0^t \mathcal{A}f(X_s)ds$$

is a martingale. A problem of this type is classified as *the martingale problem.* The above formula is a direct consequence of an application of the Itô formula onto f, and so $\mathcal{A}$ is the infinitesimal generator of X_t. In contrast, when dealing with a reflected diffusion, we instead have what is known as *the submartingale problem*, where the aim is to find an operator $\mathcal{L}$ such that

$$f(X_t) - \int_0^t \mathcal{L}f(X_s)ds$$

is a submartingale — see, for example, [KR14].

To find some notion of a generator for the process in (5.7) we shall apply the

Itô formula for some $f \in C^2(M)$,

$$\begin{aligned} df(X_t) &= g(\nabla f(X_t), dX_t) + \frac{1}{2}\mathrm{Hess}^f(dX_t, dX_t), \\ &= g(\nabla f(X_t), \Xi(X_t)dB_t) - \frac{1}{2}g(\nabla\phi(X_t), \nabla f(X_t))dt \\ &\quad + g(\nu(X_t), \nabla f(X_t))dL_t + \frac{1}{2}\sum_{i=1}^{n}\Xi_t^2(e_i)f(X_t)dt, \end{aligned}$$

where $e_1, ..., e_n$ is a basis in $\mathbb{R}^n$. The final term in this equation is none other than Bochner's horizontal Laplacian, (2.2). Applying Lemma 2.3.1 allows us to simplify the right most term, yielding the more familiar form

$$\begin{aligned} df(X_t) &= g(\nabla f(X_t), \Xi(X_t)dB_t) - \frac{1}{2}g(\nabla\phi(X_t), \nabla f(X_t))dt \\ &\quad + g(\nu(X_t), \nabla f(X_t))dL_t + \frac{1}{2}\Delta_M f(X_t)dt. \end{aligned}$$

Further simplification of the right hand side can be achieved by noting that $\frac{1}{2}\Delta_M f(x) - \frac{1}{2}g(\nabla\phi(X_t), \nabla f(X_t))$ is the infinitesimal generator of the unreflected process (5.4).

Applying the usual integral rules and taking expectation of both sides with respect to the law of X_t given that we start $X_0 = 0$, then

$$P_t f(x) = f(x) + \int_0^t \mathbb{E}[\mathcal{A}f(X_s)]ds + \mathbb{E}\left[\int_0^t g(\nu(X_s), \nabla f(X_s))dL_s\right].$$

The main complication that arises from this formulation of the semigroup is the local time term. One typical way to bypass this term is to impose a further condition on f; we can re-write $g(\nu, \nabla f) = \frac{\partial f}{\partial \nu} = df(\nu)$, the directional derivative of f in the normal direction ν. We now impose the condition $df(\nu)|_{\partial M} = 0$ to eliminate the local time integral, since L only increases on ∂M. Under this newly restricted function space $\mathcal{D} := \{f \in C_0^2(M) : df(\nu)|_{\partial M} = 0\}$, the semigroup simplifies to

$$P_t f(x) = f(x) + \int_0^t \mathbb{E}[\mathcal{A}f(X_s)]ds.$$

At this point, we derive the 'infinitesimal generator' of the reflected process,

$$\frac{d}{dt}P_t f(x)\bigg|_{t=0} = \mathcal{A}f(x).$$

To conclude, we say that the infinitesimal generator of the reflected process is equivalent to the infinitesimal generator of the unreflected process (5.3) when restricted to the function space $\mathcal{D}$. Now that we have obtained an operator, the next step is to verify that $\mathbb{E}_{\mu_\phi}[\mathcal{A}f] = 0$. This is equivalent to showing that μ_ϕ is an invariant measure for the unreflected process (5.4).

Remark. For a general class of functions, it is difficult to find the infinitesimal generator, let alone the invariant distribution. In [KR14], Kang and Ramanan redefine the notion of stationarity for the submartingale problem. They define that a probability measure π on closed space $\bar{G}$ is a stationary measure if

$$\int_0^t \mathcal{L}f(x)\pi(dx) \leq 0$$

for measures π such that $\pi(\partial G) = 0$. We differ in our approach, however, since we restrict the support of the operator $\mathcal{L}$ further. Moreover, they do not assume that the boundary is C^∞, which is an important assumption that we rely upon.

Using Green's identity on manifold [GHL90],

$$\begin{aligned}
\mathbb{E}_{\mu_\phi}[\Delta_M f] &= \int_M \Delta_M f d\mu_\phi, \\
&= \frac{1}{C_\phi}\int_M e^{-\phi}\Delta_M f d\text{vol}, \\
&= -\frac{1}{C_\phi}\int_M g(\nabla e^{-\phi}, \nabla f)d\text{vol} + \frac{1}{C_\phi}\int_{\partial M} e^{-\phi} df(\nu)\iota_\nu d\text{vol},
\end{aligned}$$

in which $\iota_\nu d$vol is the induced volume form on ∂M. Then, because $f \in \mathcal{D}$, $df(\nu) = 0$ and so the right hand most integral vanishes,; leaving us with

$$\mathbb{E}_{\mu_\phi}[\Delta_M f] = -\frac{1}{C_\phi}\int_M g(\nabla e^{-\phi}, \nabla f)d\text{vol},$$

$$= \int_M g(\nabla\phi, \nabla f) \frac{e^{-\phi}}{C_\phi} d\mathrm{vol},$$
$$= \mathbb{E}_{\mu_\phi}[g(\nabla\phi, \nabla f)].$$

Whence, we have confirmed that $\mathbb{E}_{\mu_\phi}[\mathcal{A}f] = 0$ for all functions $f \in \mathcal{D}$.

We summarise what we have established in the following theorem:

Theorem 5.2.1. *In the set $f \in \mathcal{D} := \{f \in C^2(M) : df(\nu)(x) = 0,\ \forall x \in \partial M\}$, the infinitesimal generator of the process* (5.7) *coincides with that of the ordinary process* (5.4)*. Moreover, the invariant distribution of* (5.7) *is μ_ϕ given by* (5.1)*.*

5.3 Coupled Diffusions

The next major step in our framework is to construct a coupled diffusion. This idea is paramount to proving that the Stein method we present is correct and provides finite bounds on the derivatives of our solution to the Stein equation. The coupling constructed is similar to, but not quite, a Kendall coupling, as we shall preserve the vector field after the parallel transport and not mirror it.

The coupling that we shall use will be as follows: We initiate the coupling at the point $(x, y) \in M \times M$ and define the tangent vector v_0 to be the vector on T_xM such that $\dot{\gamma}(0) = v_0$ in which γ is the unique geodesic connecting x and y; by this we mean $\gamma(0) = x$ and $\gamma(\rho(x, y)) = y$ with $\rho(x, y) = |v_0| = \sqrt{g(v_0, v_0)}$. A more concrete way to define v_0 is via the exponential mapping from x to y, $y = \exp_x(v_0)$. If one assumes that y is not a cut point of x, v_0 is unique, hence exp is an injection. This implies that the exponential map is invertible, $v_0 = \exp_x^{-1}(y)$ and moreover, the mapping $(x, v_0) \mapsto (x, \exp_x(v_0))$ is a local diffeomorphism. We may extend this notion to determine a process on the tangent bundle with which we can use the exponential map to propagate our original process X_t in the direction of some V_t to Y_t. Explicitly, we define the pair $(X_t, V_t) \in TM$ and $Y_t = \exp_{X_t}(V_t)$. The coupling process is governed by the following set of SDEs:

$$\begin{aligned} dX_t &= \Xi(X_t) \circ dB_t - \frac{1}{2}\nabla\phi(X_t)dt + \nu(X_t)dL_t^X, \quad X_0 = x, \\ dY_t &= \Upsilon(Y_t) \circ dB'_t - \frac{1}{2}\nabla\phi(Y_t)dt + \nu(Y_t)dL_t^Y, \quad Y_0 = y, \\ d\Xi_t &= H_\Xi \circ dX_t, \quad \Xi_0 = \xi, \\ d\Upsilon_t &= H_\Upsilon \circ dY_t, \quad \Upsilon_0 = \eta, \\ dB'_t &= \big(\Upsilon_t^{-1}\Pi_{X_t,V_t}\Xi_t\big)dB_t. \end{aligned} \tag{5.9}$$

In the coupling above, as with Ξ and ξ for X_t, Υ and η are the lift of Y_t to $\mathcal{O}(M)$. The new B'_t process is another Brownian motion on $\mathbb{R}^n$, however, this is dependent upon the original Brownian motion B_t. In essence, we are parallel translating the Brownian motion $\Xi_t \circ dB_t$ along the geodesic connecting X_t and Y_t to create the new Brownian motion $\Upsilon_t \circ dB'_t$. One may write (with an abuse of notation) that $\Upsilon_t dB'_t = \Pi_{X_t,V_t}\Xi_t dB_t$ which makes this fact more evident.

We write $\rho(X_t, Y_t)$ for the length of the intervening unit-speed geodesic γ_t between X_t and $Y_t = \exp_{X_t}(V_t)$. Therefore $\gamma_t(s) = \exp_{X_t}(s\frac{V_t}{|V_t|})$ so that $\gamma_t(\rho(X_t, Y_t)) = Y_t$. For convex manifolds, V_t will always be unique since the cut locus at every point in M is empty.

Let $u_1, u_2, ..., u_n$ be an orthonormal basis in $\mathbb{R}^n$ such that $\Xi_t u_1 = \dot\gamma_t(0) = V_t/|V_t|$, and further define a new basis $v_i = (\Upsilon_t^{-1}\Pi_{X_t,V_t}\Xi_t)u_i$ for $i = 1, ..., n$. We observe that $\Upsilon_t v_1 = \Pi_{X_t,V_t}\dot\gamma_t(0) = \dot\gamma_t(\rho(X_t, Y_t))$. For notational convenience, we write $\nabla_x\rho(x, y)$ as the gradient of ρ with respect to the first n arguments, and ∇_y with respect to the final n arguments. We remind the reader of the canonical projection $\pi : \mathcal{O}(M) \to M$ and denote by $\tilde\rho(\xi, \eta)$, for any $\xi, \eta \in \mathcal{O}(M)$, the projected distance $\rho(\pi\xi, \pi\eta)$ to $\mathcal{O}(M)$. Then, the Itô formula for $\rho(X_t, Y_t)$ is

$$\begin{aligned} d\rho(X_t, Y_t) = &\sum_{i=1}^{n}(\Xi_t u_i)\tilde\rho(X_t, Y_t)g(u_i, dB_t) + (\Upsilon_t v_i)\tilde\rho(X_t, Y_t)g(v_i, dB'_t) \\ &- \frac{1}{2}\Big(g(\nabla_x\rho(X_t, Y_t), \nabla\phi(X_t)) + g(\nabla_y\rho(X_t, Y_t), \nabla\phi(Y_t))\Big)dt \end{aligned}$$

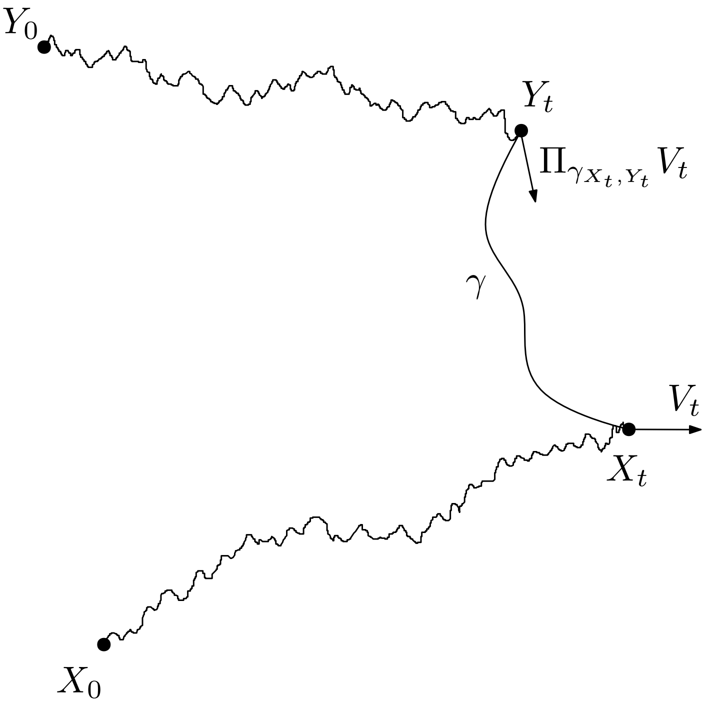

Figure 5.2: Paths of the diffusions X_t and Y_t alongside the parallel transport of V_t.

$$\begin{aligned}
&+ g(\nabla_x \rho(X_t, Y_t), \nu(X_t))dL^X(t) + g(\nabla_y \rho(X_t), \nu(Y_t))dL_t^Y \\
&+ \frac{1}{2}\sum_{i=1}^{n}(\Xi_t u_i + \Upsilon_t v_i)^2 \tilde{\rho}(X_t, Y_t)dt,
\end{aligned}$$

which is derived from Theorem 3 of [Ken86] with orthonormal basis $\{(u_1, 0), (u_2, 0), ..., (u_n, 0), (0, v_1), (0, v_2), ..., (0, v_n)\}$ on $\mathbb{R}^n \oplus \mathbb{R}^n$. The drift and local time terms are extensions from $\mathbb{R}^n$. Note that there are no further terms involving local time since L^X and L^Y are both finite variation processes. We then draw attention to our carefully chosen bases $\{u_i\}$ and $\{v_i\}$, where the choice is made in such a way that ρ only changes in the u_1 and v_1 directions for the x and y variables respectively. We make the simplification

$$\begin{aligned}
&\sum_{i=1}^{n}(\Xi_t u_i)\tilde{\rho}(X_t, Y_t)g(u_i, dB_t) + (\Upsilon_t v_i)\tilde{\rho}(X_t, Y_t)g(v_i, dB'_t) \\
&\quad = \dot{\gamma}_t(0)\rho(X_t, Y_t)g(u_1, dB_t) + \dot{\gamma}_t(\rho(X_t, Y_t))\rho(X_t, Y_t)g(v_1, dB'_t), \\
&\quad = \dot{\gamma}_t(0)\rho(X_t, Y_t)g(\dot{\gamma}_t(0), \Xi_t dB_t) + \dot{\gamma}_t(\rho(X_t, Y_t))\rho(X_t, Y_t)g(\dot{\gamma}_t(\rho(X_t, Y_t)), \Upsilon_t dB'_t),
\end{aligned}$$

$$= \dot{\gamma}_t(\rho(X_t, Y_t))\rho(X_t, Y_t)(-g(\dot{\gamma}_t(0), \Xi_t dB_t) + g(\dot{\gamma}_t(\rho(X_t, Y_t)), \Upsilon_t dB'_t),$$

since the horizontal lift to $\mathcal{O}(M)$ is an isometry and $\dot{\gamma}_t(0)\rho(X_t, Y_t) = -\dot{\gamma}_t(\rho(X_t, Y_t))\rho(X_t, Y_t)$. The latter can be seen by utilising $\nabla_x \rho(X_t, Y_t) = -\dot{\gamma}_t(0)$ and $\nabla_y \rho(X_t, Y_t) = \dot{\gamma}_t(\rho(X_t, Y_t))$. Whence, the differential of ρ is now written as

$$\begin{aligned} d\rho(X_t, Y_t) = {} & \dot{\gamma}_t(\rho(X_t, Y_t))\rho(X_t, Y_t)(-g(\dot{\gamma}_t(0), \Xi_t dB_t) + g(\dot{\gamma}_t(\rho(X_t, Y_t)), \Upsilon_t dB'_t) \\ & - \frac{1}{2}\Big(g(\nabla_x \rho(X_t, Y_t), \nabla\phi(X_t)) + g(\nabla_y \rho(X_t, Y_t), \nabla\phi(Y_t))\Big)dt \\ & + g(\nabla_x \rho(X_t, Y_t), \nu(X_t))dL^X(t) + g(\nabla_y \rho(X_t), \nu(Y_t))dL_t^Y \\ & + \frac{1}{2}\sum_{i=1}^{n}(\Xi_t u_i + \Upsilon_t v_i)^2 \tilde{\rho}(X_t, Y_t)dt. \end{aligned} \tag{5.10}$$

We begin by simplifying the first bracketed term:

$$\begin{aligned} -g(\dot{\gamma}_t(0), \Xi_t dB_t) + g(\dot{\gamma}_t(\rho(X_t, Y_t)), \Upsilon_t dB'_t) & \\ & = -g(\dot{\gamma}_t(0), \Xi_t dB_t) + g(\Pi_{X_t,V_t}\dot{\gamma}_t(0), \Pi_{X_t,V_t}\Xi_t dB_t), \\ & = -g(\dot{\gamma}_t(0), \Xi_t dB_t) + g(\dot{\gamma}_t(0), \Pi^{-1}_{X_t,V_t}\Pi_{X_t,V_t}\Xi_t dB_t), \\ & = 0. \end{aligned} \tag{5.11}$$

For the dL_t terms, we again exploit $\nabla_x \rho(X_t, Y_t) = -\dot{\gamma}_t(0)$. Fix Y_t at any given point on the manifold. Then, $\dot{\gamma}(0)$ is always pointing outward from the manifold. Since the boundary ∂M is convex, then it is always the case that $g(\nu(X_t), -\dot{\gamma}_t(0)) \leq 0$ (see Figure 5.3) with equality when the geodesic between X_t and Y_t lies on ∂M. Therefore, we have the upper bound

$$g(\nabla_x \rho(X_t, Y_t), \nu(X_t))dL_t^X + g(\nabla_y \rho(X_t), \nu(Y_t))dL_t^Y \leq 0. \tag{5.12}$$

For the drift term arising from ϕ, we note that

$$\frac{\partial}{\partial s} g(\nabla\phi(\gamma_t(s)), \dot{\gamma}_t(s)) = g(D_{\dot{\gamma}_t(s)}\nabla\phi(\gamma_t(s)), \dot{\gamma}_t(s)) + g(\nabla(\gamma_t(s)), D_{\dot{\gamma}_t(s)}\dot{\gamma}_t(s)),$$

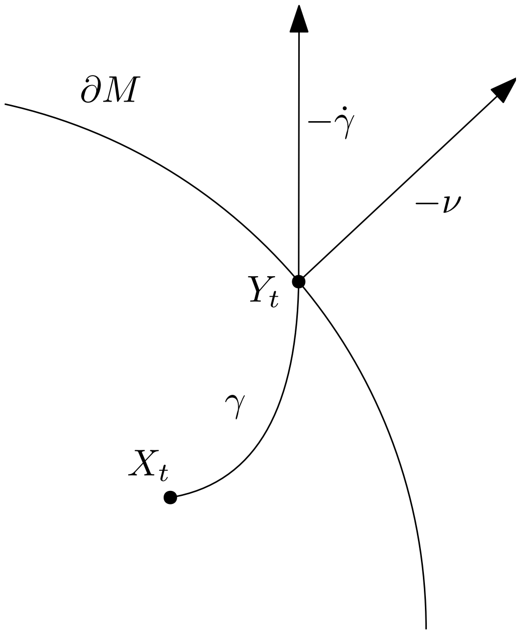

Figure 5.3: Diagram of $X_t, Y_t \in M$ with normal vector $\nu(X_t)$ and outward vector $\nabla_x \rho = -\dot{\gamma}_t(0)$.

$$\begin{aligned}&= g(D_{\dot{\gamma}_t(s)} \nabla \phi(\gamma_t(s)), \dot{\gamma}_t(s)),\\ &= \mathrm{Hess}^{\phi}(\dot{\gamma}_t(s), \dot{\gamma}_t(s)).\end{aligned}$$

It is then evident that, by integrating the above,

$$\int_0^{\rho(X_t,Y_t)} \mathrm{Hess}^{\phi}(\dot{\gamma}_t(s), \dot{\gamma}_t(s)) ds = g(\nabla \phi(Y_t), \dot{\gamma}_t(\rho(X_t, Y_t))) - g(\nabla \phi(X_t), \dot{\gamma}_t(0)), \tag{5.13}$$

which is precisely twice the negative of the first two dt terms in (5.10).

For the final term in (5.10), we apply similar reasoning as in the proof of Theorem 4 in [Ken86]. Denote by J the Jacobi vector field along the geodesic γ_t with components J_t^i, whose boundary values satisfy $J_t^i(0) = \Xi_t u_i$ and $J_t^i(\rho(X_t, Y_t)) = \Upsilon_t v_i$ for $i = 1, ..., n$. For more information on Jacobi vector fields, refer to Appendix A. Also denote $\gamma_{t,\epsilon}$ as a variation of γ_t with similar fixed boundary conditions $\gamma_{t,\epsilon}(0) = X_t$ and $\gamma_{t,\epsilon}(\rho(X_t, Y_t)) = Y_t$ for some suitable ϵ in an interval $(-a, b)$ containing 0. We represent the length of the variation $\gamma_{t,\epsilon}$ by $\rho(\epsilon)$. When $\epsilon = 0$

we recover the geodesic γ_t. Then, by the second variation formula (A.2)

$$\begin{aligned}\sum_{i=1}^{n}(\Xi u_i + \Upsilon v_i)^2\tilde{\rho}(X_t, Y_t) &= \frac{\partial^2}{\partial\epsilon^2}\rho(\epsilon)\bigg|_{\epsilon=0} = I(J_t, J_t) \\ &= \int_0^{\rho(X_t,Y_t)} |D_{\dot{\gamma}_t(s)}J_t(s)|^2 - g(R(J_t(s), \dot{\gamma}_t(s))\dot{\gamma}_t(s), J_t(s))ds \\ &= \int_0^{\rho(X_t,Y_t)} \sum_{i=1}^{n} |D_{\dot{\gamma}_t(s)}J_t^i(s)|^2 - g(R(J_t^i(s), \dot{\gamma}_t(s))\dot{\gamma}_t(s), J_t^i(s))ds,\end{aligned}$$

due to the linearity of the covariant derivative. By definition, Jacobi vector fields minimize the index form under the conditions that for any $U^i \in TM$, $U^i(0) = \Xi_t u_i$ and $U^i(\rho(X_t, Y_t)) = \Upsilon_t v_i$ — see Lemma 1.21 of [CE08]. From this we shall bound the index form $I(J_t^i, J_t^i)$ above by $I(V_t^i, V_t^i)$ where $V_t^i(s) := (\Pi_{X_t, sV_t/|V_t|}\Xi_t)u_i$. Therefore, we have the bound

$$\sum_{i=1}^{n}(\Xi u_i + \Upsilon v_i)^2\rho(X_t, Y_t) \leq \int_0^{\rho(X_t,Y_t)} \sum_{i=1}^{n} |D_{\dot{\gamma}_t(s)}V_t^i(s)|^2 - g(R(V_t^i(s), \dot{\gamma}_t(s))\dot{\gamma}_t(s), V_t^i(s))ds.$$

Since $V_t^i(s)$ was defined to be transported in parallel from X_t to Y_t, the first term involving $D_{\dot{\gamma}_t}V_t^i$ vanishes, leaving us with Riemannian curvature. Moreover, because $V_t^i(s)$ produces an orthonormal basis on $T_{\gamma_t(s)}M$, the Riemannian curvature tensor can be contracted. This leaves us with an integral involving the Ricci curvature

$$\sum_{i=1}^{n}(\Xi u_i + \Upsilon v_i)^2\tilde{\rho}(X_t, Y_t) \leq -\int_0^{\rho(X_t,Y_t)} \mathrm{Ric}(\dot{\gamma}_t(s), \dot{\gamma}_t(s))ds. \tag{5.14}$$

Combining the bounds and equalities in (5.11), (5.12), (5.13), and (5.14) into the Itô formula for the distance function (5.10) we obtain the following, crucial bound on the distance function,

$$d\rho(X_t, Y_t) \leq -\frac{1}{2}\int_0^{\rho(X_t,Y_t)} \mathrm{Ric}(\dot{\gamma}_t(s), \dot{\gamma}_t(s)) + \mathrm{Hess}^{\phi}(\dot{\gamma}_t(s), \dot{\gamma}_t(s))dsdt. \tag{5.15}$$

With the upper bound (5.15), we have the following theorem which is paramount to our construction of the Stein method on manifolds with boundary.

Theorem 5.3.1. *Let M be a convex manifold with boundary. Let (X_t, Y_t) be the coupled diffusions as defined in* (5.9) *which is initiated at the point (x, y). Assume that the Bakry-Émery Ricci curvature is bounded below as in* (5.6) *for some $\kappa > 0$. Then, for any $q \geq 1$,*

$$\rho(X_t, Y_t)^q \leq \rho(x, y)^q e^{-q\kappa t}.$$

Proof. For $q = 1$, we may immediately apply the sufficient condition (5.6) onto (5.15) to obtain

$$\begin{aligned} d\rho(X_t, Y_t) &\leq -\kappa \int_0^{\rho(X_t,Y_t)} g(\dot{\gamma}_t(s), \dot{\gamma}_t(s)) ds dt \\ &= -\kappa \int_0^{\rho(X_t,Y_t)} ds dt \\ &= -\kappa \rho(X_t, Y_t) dt. \end{aligned} \tag{5.16}$$

Then integrating and using the initial condition $\rho(X_0, Y_0) = \rho(x, y)$ we obtain the desired result,

$$\rho(X_t, Y_t) \leq \rho(x, y) e^{-\kappa t}.$$

For $q > 1$, we apply the Itô formula to the function $f(\rho) = \rho^q$,

$$\begin{aligned} d\rho(X_t, Y_t)^q &= q\rho(X_t, Y_t)^{q-1} d\rho(X_t, Y_t) + \frac{q(q-1)}{2} \rho(X_t, Y_t) d[\rho(X, Y)]_t \\ &= q\rho(X_t, Y_t)^{q-1} d\rho(X_t, Y_t) \\ &\leq -\kappa q \rho(X_t, Y_t)^q dt \end{aligned}$$

and applying integration again with the required initial condition yields the result. □

This theorem tells us that X_t and Y_t will eventually meet at the same point, say X, almost surely. Moreover, from the previous section, we have shown that this X is distributed according to the measure (5.1) and so we can conclude that

Y_t's invariant distribution is also X. If X_t and Y_t meet before infinity, then the coupling terminates and $\{Y_t\}$ becomes an identical copy of X_t.

5.4 The Stein Equation

The remainder of the chapter will now concern itself with the study of the Stein equation and solution. In this section, we shall derive the solution and formulate bounds on the solution and its first derivative.

The Stein equation is a second order PDE of the form

$$\mathcal{A}f_h(x) = h(x) - \mathbb{E}[h(X)] \tag{5.17}$$

in which $\mathcal{A}$ is as in Theorem 5.2.1 and X is the invariant distribution — $X \sim \mu_\phi$.

Lemma 5.4.1. *Define the space of test functions $\mathcal{H} = \{h : M \to \mathbb{R} : C_0(h) < \infty\}$. Define $\{X_t\}_{t\in\mathbb{R}^+}$ to be the diffusion according to the SDE* (5.7) *with generator $\mathcal{A}$ and invariant distribution $X \sim \mu_\phi$. Moreover, we assume that $\mathbb{E}[\rho(x, X)] < \infty$ for any point $x \in M$. Then, the solution to the Stein equation* (5.17) *is*

$$f_h(x) = \int_0^\infty \mathbb{E}[h(X)] - \mathbb{E}_x[h(X_t)]dt.$$

Furthermore, f_h is well defined and has Lipschitz constant $C_0(f_h) \leq C_0(h)/\kappa$.

Proof. Define $\{(X_t, Y_t)\}_{t\in\mathbb{R}^+}$ to be the coupled diffusions as constructed in (5.9). First, we show that f_h is well defined:

$$\begin{aligned}
|f_h(x)| &= \left|\int_0^\infty \mathbb{E}[h(X)] - \mathbb{E}_x[h(X_t)]dt\right| \\
&= \left|\int_0^\infty \int_M \mathbb{E}_y[h(Y_t)] - \mathbb{E}_x[h(X_t)]d\mu_\phi(y)dt\right| \\
&\leq C_0(h)\int_0^\infty \int_M \mathbb{E}_{x,y}[\rho(X_t, Y_t)]d\mu_\phi(y)dt \\
&\leq C_0(h)\int_0^\infty \int_M \rho(x, y)e^{-\kappa t}\mu_\phi(y)dt
\end{aligned}$$

$$\begin{aligned}
&= C_0(h)\mathbb{E}[\rho(x, X)] \int_0^\infty e^{-\kappa t} dt \\
&= \frac{C_0(h)}{\kappa}\mathbb{E}[\rho(x, X)] < \infty.
\end{aligned}$$

Here, the second equality is due to the fact that X is the invariant measure of Y_t, the first inequality is due to the definition of Lipschitz continuity of h, and the second inequality is the application of Theorem 5.3.1.

To prove the condition on the Lipschitz constant of f_h, we note that

$$\begin{aligned}
|f_h(x) - f_h(y)| &= \left| \int_0^\infty \mathbb{E}_y[h(Y_t)] - \mathbb{E}_x[h(X_t)] dt \right|, \\
&\leq C_0(h) \int_0^\infty \mathbb{E}_{x,y}[\rho(X_t, Y_t)] dt, \\
&\leq C_0(h) \int_0^\infty \rho(x, y) e^{-\kappa t} dt, \\
&= \frac{C_0(h)}{\kappa} \rho(x, y),
\end{aligned}$$

which implies that $C_0(f_h) \leq C_0(h)/\kappa$. This also shows that $df_h(x)(u)$ is bounded in some general direction $u \in T_x M$.

Lastly, we prove that f_h is indeed the solution to the Stein equation: in order to show this, we must prove that

$$f_h(x) = \int_0^\infty \mathbb{E}[h(X)] - \mathbb{E}_x[h(X_t)] dt$$

satisfies the differential equation

$$h(x) - \mathbb{E}[h(X)] = \mathcal{A} f_h(x).$$

First, assuming that f_h satisfies the condition $df_h(x)(\nu) = 0$ for all $(x, \nu) \in N(\partial M)$, define $P_t h(x) = \mathbb{E}_x[h(X_t)]$ as the semigroup of $\{X_t\}_{t \in \mathbb{R}^+}$. Then $\mathcal{A}$ is the infinitesimal generator of X_t and $\frac{d}{dt} P_t|_{t=0} = \mathcal{A}$. Then, by Proposition 1.5

of [EK09],

$$\begin{aligned} h(x) - P_t h(x) &= -\mathcal{A}\int_0^t P_s h(x) ds \\ &= \mathcal{A}\int_0^t \mathbb{E}[h(X)] - P_s h(x) ds, \end{aligned} \tag{5.18}$$

since $\mathbb{E}[h(X)]$ is a constant. On examination of the left hand side of (5.18), we note

$$\begin{aligned} |h(x) - \mathbb{E}[h(X)] - (h(x) - P_t h(x))| &= |P_t h(x) - \mathbb{E}[h(X)]| \\ &= \left| \int_M \mathbb{E}_x[h(X_t)] - \mathbb{E}_y[h(Y_t)] d\mu_\phi(y) \right| \\ &\leq C_0(h) \int_M \mathbb{E}_{x,y}[\rho(X_t, Y_t)] d\mu_\phi(y) \\ &\leq C_0(h) \int_M \rho(x,y) e^{-\kappa t} d\mu_\phi(y) \\ &= C_0(h) \mathbb{E}[\rho(x,X)] e^{-\kappa t} < \infty, \end{aligned}$$

due to X being the invariant measure of $\{Y_t\}_{t\in\mathbb{R}^+}$. We then conclude that, pointwise,

$$\lim_{t\to\infty} h(x) - P_t h(x) = h(x) - \mathbb{E}[h(X)].$$

For the right hand side of (5.18), since f_h is well defined, as shown above, and $\mathcal{A}$ is closed by [EK09, Corollary 1.6], we may apply the dominated convergence theorem:

$$\lim_{t\to\infty} \mathcal{A}\int_0^t \mathbb{E}[h(X)] - P_s h(x) ds = \mathcal{A}\int_0^\infty \mathbb{E}[h(X)] - \mathbb{E}_x[h(X_t)] dt.$$

To conclude, f_h is the solution to the Stein equation. □

Since the encompassing manifold N can be embedded within some higher dimensional Euclidean space via Whitney's Embedding Theorem, so can the manifold with boundary M. As a consequence, we can make use of stochastic analysis on smooth domains embedded within $\mathbb{R}^n$. Particularly, [And11, Corollary 2.9] tells us that the Neumann condition on the semigroup $dP_t h(x)(\nu) = 0$ is automatically satisfied along the boundary. The main use of this result is to show that the form of f_h satisfies the Neumann boundary condition that we require for the infinitesimal generator:

$$df_h(x)(\nu) = \int_0^\infty dP_t h(x)(\nu)dt = 0$$

since $h \in \mathcal{H} \subset C(M)$.

Since $\mathcal{A}$ is a second order differential operator, it is imperative to also check that $\Delta_M f_h$ is well defined. To do this, we shall verify that $\|D^2 f_h\|_{\text{op}} < \infty$. This, however, is not an easy task, and requires the use of damped stochastic parallel translation.

5.5 Bounding the Second Derivative

5.5.1 Weitzenböck Formula

We dedicate this subsection to the introduction of the adjoint operator of d, the Hodge–de Rahm Laplacian, and the Weitzenböck formula.

Let α and β be two differential forms of the same degree q with compact support. We define the L^2 inner product between α and β as

$$(\alpha, \beta) = \int_M \langle \alpha, \beta \rangle_x d\text{vol}(x), \tag{5.19}$$

where $\langle \cdot, \cdot \rangle$ is the induced inner product on $(T^*M)^q = T^*M \otimes ... \otimes T^*M$ q times. Let $\delta : \Gamma(\Lambda^p M) \to \Gamma(\Lambda^{p-1} M)$ be defined as the formal adjoint of d with respect to the inner product (5.19). The existence and uniqueness of δ is guaranteed by the Riesz representation theorem. Explicitly, for $\omega \in \Lambda^q(M)$ and $\tau \in \Lambda^{q-1}(M)$

$$(d\tau, \omega) = (\tau, \delta\omega).$$

The adjoint δ lowers the degree of the form by 1, in contrast with the exterior derivative d which increases it by 1. As with the the exterior derivative, it is the case that $\delta^2 = 0$. δ is sometimes called the divergence operator. For a function $f \in C^1(M)$, $\delta f = 0$. See [Jos08, Section 3.3] for a more rigorous treatment.

We now define the Hodge-de Rahm Laplacian,

$$\Box_M = -(d\delta + \delta d).$$

One major difference between Δ_M and $\Box_M$ that gives $\Box_M$ a more geometrical significance is that it commutes with d,

$$\Box_M d = -(d\delta + \delta d)d = -d\delta d = d\Box_M.$$

The Hodge-de Rahm Laplacian can also be written as the operator $\Box_M = \mathrm{Tr}D^2$, where D is the connection on TM. Therefore on functions, $\Box_M$ and Δ_M coincide. On forms, however, they differ by a linear transformation on $\Gamma(\Lambda^\bullet M)$ controlled by the curvature tensor. This well known formula is known as the Weitzenböck formula (see [Hsu02]).

Lemma 5.5.1. *Let f be a function on M and df on T^*M, then*

$$\Box_M df = d\Delta_M f + (df)(\mathrm{Ric}^\#).$$

The # notation in the Ricci tensor is known as the sharp musical isomorphism, the role of which transforms the 2-form to a tensor of type (1,1) — $\mathrm{Ric}^\# : T^*M \to T^*M$ or $\mathrm{Ric}^\# : TM \to TM$. In Einstein summation notation, this is written as $(\mathrm{Ric}^\#)_i^j = g^{jk}\mathrm{Ric}_{ik}$. With the inner product, for vector fields $U, V \in T_xM$, $g(\mathrm{Ric}^\#(U), V) = \mathrm{Ric}(U, V)$.

For use in analysis, we also require the following result [Tho20]:

Lemma 5.5.2. *For any vector fields X and Z and function f, the following is* *true*

$$d(Z(f))(X) = (D_Z df)(X) + df(D_X Z).$$

5.5.2 Damped Stochastic Parallel Translation

We first define the process $\{W_t\}_{t\in\mathbb{R}^+}$, $W_t : T_{X_0}M \to T_{X_t}M$, the damped stochastic parallel translation of the process $\{X_t\}_{t\in\mathbb{R}^+}$. It is the solution of the Stochastic Covariant Differential Equation (SCDE) given by

$$DW_t = -\frac{1}{2}(\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#}(W_t)dt + D\nu(W_t)dL_t, \quad W_0 = \mathrm{Id}. \tag{5.20}$$

Damped stochastic parallel translations have been primarily used to generate Bismut–Elworthy–Li type formulae for the semigroup P_t in which one can exchange d and P_t in a careful manner. This has been done with drift and no local time in [Tho20] and local time and no drift in [AL17].

Theorem 5.5.3. *Suppose the sufficient condition* (5.6) *is satisfied for the pair M and ϕ. Let W_t be the solution of the CSDE* (5.20)*. Then the process $dP_t h(W_t)$ is a martingale and therefore*

$$dP_t h(u) = \mathbb{E}[(dh)(W_t(u))].$$

Proof. We begin by applying the Itô formula for 1-forms (see [Li92, Equation (1.4)]) to the function $dP_{t-s}h(W_s) = g(\nabla P_{t-s}h, W_s)$:

$$\begin{aligned} d(g(\nabla P_{t-s}h, W_s)) = {} & \nabla dP_{t-s}h(\Xi_s(X_s)dB_s, W_s) + \nabla dP_{t-s}h(\nu(X_s), W_s)dL_s \\ & + \left(\partial_s + \frac{1}{2}\mathrm{Tr}D^2 - \frac{1}{2}D_{\nabla\phi}\right)dP_{t-s}h(W_s)ds + dP_{t-s}h(DW_s). \end{aligned} \tag{5.21}$$

The final term can be written out in full using (5.20)

$$dP_{t-s}h(DW_s) = -\frac{1}{2}g(\nabla P_{t-s}h, (\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#}(W_s))ds + g(\nabla P_{t-s}h, D\nu(W_s))dL_s. \tag{5.22}$$

Using Lemma 5.5.2 we have the relation $d(\nabla\phi(f)) - dh(\mathrm{Hess}^{\phi}) = D_{\nabla\phi}df$, which can then be substituted into (5.21) alongside (5.22) to give

$$\begin{aligned} d(g(\nabla P_{t-s}h, W_s)) = {} & \nabla dP_{t-s}h(\Xi_s(X_s)dB_s, W_s) + \nabla dP_{t-s}h(\nu(X_s), W_s)dL_s \\ & + \left(\partial_s + \frac{1}{2}\mathrm{Tr}D^2\right)dP_{t-s}h(W_s)ds \\ & - \frac{1}{2}\big(d(\nabla\phi(P_{t-s}h)) - dP_{t-s}h(\mathrm{Hess}^{\phi})\big)(W_s)dt \\ & - \frac{1}{2}dP_{t-s}h(\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#}(W_s)ds + g(\nabla P_{t-s}h, D_{W_s}\nu(X_s))dL_s. \end{aligned}$$

By noting that

$$\begin{aligned} dP_{t-s}h(\mathrm{Hess}^{\phi})^{\#}(W_s) &= g(\nabla P_{t-s}h, (\mathrm{Hess}^{\phi})^{\#}(W_s)), \\ &= \mathrm{Hess}^{\phi}(\nabla P_{t-s}h, W_s) = dP_{t-s}h(\mathrm{Hess}^{\phi})(W_s), \end{aligned}$$

we can simplify the above equation further:

$$\begin{aligned} d(g(\nabla P_{t-s}h, W_s)) = {} & \nabla dP_{t-s}h(\Xi_s(X_s)dB_s, W_s) + \nabla dP_{t-s}h(\nu(X_s), W_s)dL_s \\ & + \left(\partial_s + \frac{1}{2}\Box_M\right)dP_{t-s}h(W_s)ds - \frac{1}{2}d(\nabla\phi)(P_{t-s}h)(W_s)ds \\ & - \frac{1}{2}dP_{t-s}h(\mathrm{Ric}^{\#})(W_s)ds + g(\nabla P_{t-s}h, D_{W_s}\nu(X_s))dL_s. \end{aligned}$$

Using the fact that ∂_t and $\Box_M$ commute with d,

$$\begin{aligned} d(g(\nabla P_{t-s}h, W_s)) = {} & \nabla dP_{t-s}h(\Xi_s(X_s)dB_s, W_s) + \nabla dP_{t-s}h(\nu(X_s), W_s)dL_s \\ & + d\left(\partial_s + \frac{1}{2}\Box_M - \frac{1}{2}g(\nabla\phi, \nabla)\right)P_{t-s}h(W_s)ds \\ & - \frac{1}{2}dP_{t-s}h(\mathrm{Ric}^{\#})(W_s)dt + g(\nabla P_{t-s}h, D_{W_s}\nu(X_s))dL_s. \end{aligned}$$

Applying the Weitzenböck formula in Lemma 5.5.1 cancels any remaining $\mathrm{Ric}^{\#}$ terms as well as producing the infinitesimal generator in the drift term:

$$\begin{aligned} d(g(\nabla P_{t-s}h, W_s)) &= \nabla dP_{t-s}h(\Xi_s(X_s)dB_s, W_s) \\ &\quad + d\left(\partial_s + \frac{1}{2}\Delta_M - \frac{1}{2}g(\nabla\phi, \nabla)\right)P_{t-s}h(W_s)ds \\ &\quad + \nabla dP_{t-s}h(\nu(X_s), W_s)dL_s + g(\nabla P_{t-s}h, D_{W_s}\nu(X_s))dL_s. \end{aligned}$$

The drift term will vanish since $P_{t-s}h$ is a solution of the backward Kolmogorov equation

$$\left(\partial_s + \frac{1}{2}\Delta_M - \frac{1}{2}g(\nabla\phi, \nabla)\right)f = 0,$$

leaving three final terms

$$\begin{aligned} d(g(\nabla P_{t-s}h, W_s)) &= \nabla dP_{t-s}h(\Xi_s(X_s)dB_s, W_s) \\ &\quad + \nabla dP_{t-s}h(\nu(X_s), W_s)dL_s + g(\nabla P_{t-s}h, D_{W_s}\nu(X_s))dL_s. \end{aligned} \tag{5.23}$$

Since $\nu \in \ker dP_th$, $DdP_th(\nu) = 0$, or in a more useful form $Dg(\nabla P_th, \nu) = 0$,

$$\begin{aligned} D_{W_s}g(\nabla P_{t-s}h, \nu) &= g(D_{W_s}dP_{t-s}h, \nu) + g(\nabla P_{t-s}h, D_{W_s}\nu) \\ &= \nabla dP_{t-s}h(\nu, W_s) + g(\nabla P_{t-s}h, D_{W_s}\nu) \\ &= 0. \end{aligned}$$

Therefore, the final two terms in (5.23) vanish and we are left with a single Itô integral

$$d(g(\nabla P_{t-s}h, W_s)) = \nabla dP_{t-s}h(\Xi_s(X_s)dB_s, W_s), \tag{5.24}$$

which indicates that $g(\nabla P_{t-s}h, W_s)$ is a local martingale.

Since $\mathrm{Ric}+\mathrm{Hess}^{\phi} \geq 2\kappa g$, it follows that $g(\nabla P_th, W_t)$ is bounded and is therefore a martingale on the interval $[0,t]$. Integrating (5.24) yields

$$dP_0h(W_t) = dP_th(u) + \int_0^t \nabla dP_{t-s}h(\Xi_s(X_s)dB_s, W_s).$$

The desired result then follows by applying expectation to both sides. □

Remark. It is not completely necessary that $\mathrm{Ric} + \mathrm{Hess}^\phi$ is bounded from below by a positive constant, only that $\mathrm{Ric}+\mathrm{Hess}^\phi \geq -\kappa g$ so that the process $\{X_t\}$ does not leave M.

5.5.3 Bound on $C_1(f_h)$

We use (5.8) to define the Lipschitz constant of the derivative of f_h as

$$C_1(f_h) = \sup_{x,y\in M} \frac{|(df_h(x) - \Pi_{\gamma_{x,y}} df_h(y))(u)|}{\rho(x,y)}, \tag{5.25}$$

where $u \in T_xM$ has norm 1. The exterior derivative d can be brought inside the integral of f_h via the Leibniz integral rule resulting in

$$df_h(x) = -\int_0^\infty dP_th(x)dt.$$

Using Theorem 5.5.3, we may take the exterior derivative inside of the semigroup, so

$$dP_th(x)(u) = \mathbb{E}[dh(X_t)(W_t)],$$

where W_t is the solution to the SCDE (5.20). We now write the numerator of (5.25) as

$$\begin{aligned}
|(df_h(x) - \Pi_{x,v} df_h(y))(u)| &= |(df_h(x)(u) - df_h(y)(\Pi_{x,v}^{-1}u)| \\
&= \left| \int_0^\infty d\mathbb{E}_x[h(X_t)](u) - d\mathbb{E}_y[h(Y_t)](\Pi_{X_t,V_t}^{-1}u)dt \right| \\
&\leq \int_0^\infty \mathbb{E}[|dh(X_t)(W_t(u)) - dh(Y_t)(W'_t(w))|]dt, \quad (5.26)
\end{aligned}$$

where W_t is the stochastic damped parallel translation for X_t, and W'_t is the stochastic damped parallel translation for Y_t, with $W_0(u) = u$ and $W'_0(w) = w := \Pi^{-1}_{x,v}u$. We compress the notation here and make the initial propagating vector fields u and w implied for W_t and W'_t respectively. We shall now split this integral up by adding 0,

$$\begin{aligned}\int_0^\infty \mathbb{E}[|dh(X_t)(W_t) - dh(Y_t)(W'_t)|]dt \\ &= \left| \int_0^\infty \mathbb{E}[(dh(X_t) - \Pi_{X_t,V_t} dh(Y_t))(W_t)] \right. \\ &\quad \left. + \mathbb{E}[dh(Y_t)(\Pi^{-1}_{X_t,V_t} W_t) - dh(Y_t)(W'_t)]dt \right| \\ &\leq \int_0^\infty \mathbb{E}[|(dh(X_t) - \Pi_{X_t,V_t} dh(Y_t))(W_t)|]dt \\ &\quad + \int_0^\infty \mathbb{E}[|dh(Y_t)(\Pi^{-1}_{X_t,V_t} W_t) - dh(Y_t)(W'_t)|]dt. \qquad (5.27)\end{aligned}$$

We begin by tackling the first integral of (5.27). To start, we construct an upper bound on the integrand. For numbers $p, q \geq 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$,

$$\begin{aligned}\mathbb{E}[|(dh(X_t) - \Pi_{X_t,V_t} dh(Y_t))(W_t)|] &\leq C_1(h)\mathbb{E}[\rho(X_t, Y_t)|W_t|], \\ &\leq C_1(h)\mathbb{E}[\rho(X_t, Y_t)^p]^{1/p}\mathbb{E}[|W_t|^q]^{1/q}, \\ &\leq C_1(h)\rho(x, y)e^{-\kappa t}\mathbb{E}[|W_t|^q]^{1/q}, \qquad (5.28)\end{aligned}$$

wherein we have applied Hölder's inequality in the second line and Theorem 5.3.1 in the third.

Our primary objective is to now bound $\mathbb{E}[|W_t|^q]$ above. We can achieve this by applying the Itô lemma onto the norm $|W_t|^2 = g(W_t, W_t)$; therefore

$$\begin{aligned}d|W_t|^2 &= dg(W_t, W_t) = 2g(W_t, DW_t) \\ &= 2g\left(W_t, -\frac{1}{2}(\mathrm{Ric} + \mathrm{Hess}^\phi)^\#(W_t)dt + D\nu(W_t)dL_t\right) \\ &= -(\mathrm{Ric} + \mathrm{Hess}^\phi)(W_t, W_t)dt + 2g(W_t, D\nu(W_t))dL_t.\end{aligned}$$

To obtain higher powers of $|W_t|$, we reapply the Itô formula for the function $f(x) = x^{q/2}$ in terms of the process $|W_t|^2$:

$$\begin{aligned} d|W_t|^q &= dg(W_t, W_t)^{q/2} \\ &= \frac{\partial(|W_t|^2)^{q/2}}{\partial|W_t|^2} d|W_t|^2 + \frac{1}{2}\frac{\partial^2(|W_t|^2)^{q/2}}{\partial(|W_t|^2)^2}(d|W_t|^2)^2 \\ &= \frac{q}{2}(|W_t|^2)^{q/2-1}\Big(- (\mathrm{Ric} + \mathrm{Hess}^{\phi})(W_t, W_t)dt + 2g(W_t, D\nu(W_t))dL_t\Big), \end{aligned} \tag{5.29}$$

since both t and L_t are of finite variation.

Theorem 5.5.4. *Suppose that M is a convex submanifold of N. Then we have the following bound on the norm of the damped stochastic parallel translation*

$$\mathbb{E}[|W_t|^q] \leq |u|^q e^{-\kappa q t}.$$

Proof. We begin by attempting to bound the right hand side of Equation (5.29). We recognise that the sufficient condition may be immediately applied to find an upper bound on the drift term:

$$d|W_t|^q \leq -\kappa q|W_t|^q dt + q|W_t|^{q-2}g(W_t, D\nu(W_t))dL_t.$$

We may simplify further by recognising that $-g(w, D\nu(w)) = \mathbb{II}(w, w)$, the second fundamental form. Then since M is a convex submanifold, the second fundamental form is positive semidefinite all along ∂M, and hence we obtain the much simpler bound

$$d|W_t|^q \leq -\kappa q|W_t|^q dt.$$

By applying the Gronwall inequality, we arrive at the desired result after applying expectation

$$\mathbb{E}[|W_t|^q] \leq |u|^q e^{-\kappa q t}. \tag{5.30}$$

□

We now apply Theorem 5.5.4 to the integral of (5.28) and substitute in our new bound for $\mathbb{E}[|W_t|^q]$:

$$\begin{aligned}\int_0^\infty \mathbb{E}[|(dh(X_t) - \Pi_{\gamma_{X_t,Y_t}} dh(Y_t))(W_t)|]dt &\leq C_1(h)\rho(x,y) \int_0^\infty e^{-\kappa t}\mathbb{E}[|W_t|^q]^{1/q} dt,\\ &\leq C_1(h)\rho(x,y)\int_0^\infty |u| e^{-2\kappa t} dt,\\ &= \frac{C_1(h)}{2\kappa}\rho(x,y). \end{aligned} \tag{5.31}$$

For the second integral in Inequality (5.27),

$$\int_0^\infty \mathbb{E}[|dh(Y_t)(\Pi^{-1}_{X_t,V_t}W_t) - dh(Y_t)(W'_t)|]dt \leq C_0(h)\int_0^\infty \mathbb{E}[|\Pi^{-1}_{X_t,V_t}W_t - W'_t|]dt. \tag{5.32}$$

For a bound on the latter integrand, we again rely upon applying the Itô lemma. Define as the difference in vector fields $Z_t := \Pi^{-1}_{X_t,V_t}W_t - W'_t \in T_{Y_t}M$. Then by Itô lemma, $d|Z_t|^2 = 2g(Z_t, DZ_t)$, in which the differential

$$\begin{aligned} DZ_t = D\Pi^{-1}_{X_t,V_t}W_t - DW'_t &= \Pi^{-1}_{X_t,V_t}DW_t - DW'_t,\\ &= \Pi^{-1}_{X_t,V_t}\Big(-\frac{1}{2}(\mathrm{Ric} + \mathrm{Hess}^\phi)^\#_{X_t}(W_t)dt + D\nu_{X_t}(W_t)dL^X_t\Big)\\ &\quad + \frac{1}{2}(\mathrm{Ric} + \mathrm{Hess}^\phi)^\#_{Y_t}(W'_t)dt - D\nu_{Y_t}(W'_t)dL^Y_t. \end{aligned}$$

To help, we have labelled at which point the tensor $(\mathrm{Ric}+\mathrm{Hess}^\phi)^\#$ lies, so that we may keep track of where in the tangent space the contraction happens. Progressing forward,

$$\begin{aligned} d|Z_t|^2 = -g(Z_t, \Pi^{-1}_{X_t,V_t}\big((\mathrm{Ric} + \mathrm{Hess}^\phi)^\#_{X_t}(W_t)\big) - (\mathrm{Ric} + \mathrm{Hess}^\phi)^\#_{Y_t}(W'_t))dt\\ + 2g(Z_t, \Pi^{-1}_{X_t,V_t}(D\nu_{X_t}(W_t))dL^X_t - D\nu_{Y_t}(W'_t)dL^Y_t). \end{aligned}$$

The inclusion of two different local time terms makes manipulation problematic. To continue, we apply the following approximation to the local time from [AL17]. Let X^a_t be an approximation of our original reflected diffusion X_t. We replace the local time term with another finite variation term with integrator in t. We write that

$$dX_t^a = \Xi_t(X_t^a) \circ dB_t - \frac{1}{2}\nabla\phi(X_t^a)dt + A(X_t^a)dt, \quad X_0 = x, \tag{5.33}$$

where A is a function of the closest distance from the point X_t^a to the boundary ∂M. The function A works by providing an ever increasing drift in the inward normal direction as our process X_t^a approaches the boundary, until the drift becomes almost infinite — the process never touches the boundary. The approximating process X_t^a converges in the topology of uniform convergence in probability (UCP)(see [AL17, Theorem 3.3]) as $a \to 0$. The additional drift term (which we shall now be denoting $A^a := A(X_t^a)$ for conciseness) and its derivative $|DA^a|$ are uniformly bounded from above outside some tubular neighbourhood on the boundary. In the interior, as $a \to 0$, it vanishes. By shrinking the tubular neighbourhood, together with the limit $a \to 0$, it approximates our local time terms. On the boundary, the approximation $A^a dt$ goes to $\nu(X_t)dL_t$.

The damped stochastic parallel translation of the approximation X_t^a is a trivial computation; since we only have drift terms, it takes the form

$$DW_t^a = -\frac{1}{2}(\mathrm{Ric} + \mathrm{Hess}^\phi)^\#(W_t^a)dt + DA^a(W_t^a)dt, \quad W_0^a = \mathrm{Id}.$$

Importantly, there is also convergence in damped stochastic parallel translations when contracted with one forms — see [AL17, Corollary 5.7].

We now apply approximations for both processes X_t and Y_t as described in (5.33) and label them X_t^a and Y_t^a respectively. Denote by W_t^a and $W_t^{'a}$ the approximations of the damped stochastic parallel translations W_t and W_t' respectively. Since we have convergence in W_t^a and $W_t^{'a}$, it is also the case that we have convergence for Z_t and $Z_t^a := \Pi^{-1}_{X_t^a, V_t^a} W_t^a - W_t^{'a}$. Therefore, approximating $d|Z_t|^2$ we have

$$d|Z_t^a|^2 = -g(Z_t^a, \Pi^{-1}_{X_t^a, V_t^a}\big((\mathrm{Ric} + \mathrm{Hess}^\phi)^\#_{X_t^a}(W_t^a)\big) - (\mathrm{Ric} + \mathrm{Hess}^\phi)^\#_{Y_t^a}(W_t^{'a}))dt$$

$$+ 2g(Z_t^a, \Pi^{-1}_{X_t^a,V_t^a} DA^a(W_t^a) - DA^{'a}(W_t^{'a}))dt, \tag{5.34}$$

where $A^{'a}$ is the approximation for Y_t.

For the first inner product in (5.34), we use the trick of adding 0, so

$$\begin{aligned}
-g(Z_t^a, \Pi^{-1}_{X_t^a,V_t^a}\big((\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#}_{X_t^a}(W_t)\big) - (\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#}_{Y_t^a}(W_t^{'a}))& \\
= g(Z_t^a, (\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#}_{Y_t^a}(W_t^{'a}) - (\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#}_{Y_t^a}(\Pi^{-1}_{X_t^a,V_t^a} W_t^a))& \\
+ g\bigg(Z_t^a, (\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#}_{Y_t^a}(\Pi^{-1}_{X_t^a,V_t^a} W_t^a) - \Pi^{-1}_{X_t^a,V_t^a}\big((\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#}_{X_t^a}(W_t^a)\big)\bigg)& \\
\leq -g(Z_t^a, (\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#}_{Y_t^a}(Z_t^a))& \\
+ |Z_t^a||(\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#}_{X_t^a}(W_t^a) - \Pi_{X_t^a,V_t^a}\big((\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#}_{Y_t^a}\big)(W_t^a)|& \\
\leq -2\kappa|Z_t^a|^2 + |Z_t^a||W_t^a|C_0((\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#})\rho(X_t^a, Y_t^a).& \qquad (5.35)
\end{aligned}$$

We have labelled $C_0((\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#})$ as the Lipschitz constant of the tensor $(\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#}$ which is defined in a manner similar to the function case. We assume that $C_0((\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#}) < \infty$. We have also used the fact that $(\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#}_{Y_t}(\Pi^{-1}_{X_t,V_t} W_t) = \Pi_{X_t,V_t}\big((\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#}_{Y_t}\big)(W_t)$.

For the second quantity in (5.34) which involves the approximation of local time, we use the same trick:

$$\begin{aligned}
2g(Z_t^a, \Pi^{-1}_{X_t^a,V_t^a}(DA^a(W_t^a)) - DA^{'a}(W_t^{'a}))& \\
= 2g(Z_t^a, DA^{'a}(\Pi^{-1}_{X_t^a,V_t^a} W_t^a) - DA^{'a}(W_t^{'a}))& \\
+ 2g(Z_t^a, \Pi^{-1}_{X_t^a,V_t^a}(DA^a(W_t^a)) - DA^{'a}(\Pi^{-1}_{X_t^a,V_t^a} W_t^a))& \\
\leq 2\left\|DA^a\right\|_{\mathrm{op}}|Z_t^a|^2 + 2|Z_t^a||W_t^a||DA^a - \Pi_{X_t^a,V_t^a} DA^{'a}|& \\
\leq 2\left\|DA^a\right\|_{\mathrm{op}}|Z_t^a|^2 + 2|Z_t^a||W_t^a|C_0(DA^a)\rho(X_t^a, Y_t^a),& \qquad (5.36)
\end{aligned}$$

where again $C_0(DA^a)$ is the Lipschitz constant of the tensor DA. We remind the reader that $\|DA^a\|$ is finite in the complement of the tubular neighbourhood around ∂M, see [AL17, p.13].

On combining of (5.35) and (5.36) we have

$$\begin{aligned} d|Z_t^a|^2 \leq & -2(\kappa - \|DA^a\|_{\mathrm{op}})|Z_t^a|^2 dt \\ & \quad + |Z_t^a||W_t^a|\rho(X_t^a, Y_t^a)\big(C_0((\mathrm{Ric} + \mathrm{Hess}^\phi)^\#) + C_0(DA^a)\big)dt. \end{aligned}$$

By applying the Itô lemma — considering $|Z_t^a|^q$ as a function of $|Z_t^a|^2$ — we find that the differential of $|Z_t^a|^q$ is

$$\begin{aligned} d|Z_t^a|^q &= \frac{q}{2}|Z_t^a|^{q-2} d|Z_t^a|^2, \\ &\leq -q(\kappa - \|DA^a\|_{\mathrm{op}})|Z_t^a|^q dt \\ & \quad + \frac{q}{2}(C_0((\mathrm{Ric} + \mathrm{Hess}^\phi)^\#) + C_0(DA^a))|Z_t|^{q-1}\rho(X_t^a, Y_t^a)|W_t^a|dt. \end{aligned}$$

In the case $q = 1$, we have

$$\begin{aligned} d|Z_t^a| \leq & -(\kappa - \|DA^a\|)|Z_t^a| dt \\ & \quad + \frac{1}{2}(C_0((\mathrm{Ric} + \mathrm{Hess}^\phi)^\#) + C_0(DA^a))\rho(X_t^a, Y_t^a)|W_t^a|dt. \end{aligned}$$

It follows that, by considering the integral form and applying expectations,

$$\begin{aligned} d\mathbb{E}[|Z_t^a|] \leq & -(\kappa - \|DA^a\|_{\mathrm{op}})\mathbb{E}[|Z_t^a|]dt \\ & \quad + \frac{1}{2}(C_0((\mathrm{Ric} + \mathrm{Hess}^\phi)^\#) + C_0(DA^a))\mathbb{E}[\rho(X_t^a, Y_t^a)|W_t^a|]dt. \end{aligned}$$

Now, by taking the limit $a \to 0$ and shrinking the tubular neighbourhood to length 0, we arrive on a bound for the real process $|Z_t|$:

$$\begin{aligned} d\mathbb{E}[|Z_t|] \leq & -(\kappa - \|D\nu\|_{\mathrm{op}})\mathbb{E}[|Z_t|]dt \\ & \quad + \frac{1}{2}(C_0((\mathrm{Ric} + \mathrm{Hess}^\phi)^\#) + C_0(D\nu))\mathbb{E}[\rho(X_t, Y_t)|W_t|]dt. \end{aligned}$$

Applying both Theorem 5.3.1 and 5.5.4 yields

$$d\mathbb{E}[|Z_t|] \leq -(\kappa - \|D\nu\|_{\mathrm{op}})\mathbb{E}[|Z_t|]dt + \frac{1}{2}(C_0((\mathrm{Ric}+\mathrm{Hess}^{\phi})^{\#}) + C_0(D\nu))e^{-2\kappa t}\rho(x,y)dt.$$

Here we have subtly used the fact that the initial propagation vector $u = W_0(u)$ has norm 1. By applying the Itô formula on $e^{(\kappa-\|D\nu\|_{\mathrm{op}})t}\mathbb{E}[|Z_t|]$, $\mathbb{E}[|Z_t|]$ can be directly bounded

$$\begin{aligned} d(e^{(\kappa-\|D\nu\|_{\mathrm{op}})t}\mathbb{E}[|Z_t|]) &= e^{(\kappa-\|D\nu\|_{\mathrm{op}})t}d\mathbb{E}[|Z_t|] + (\kappa - \|D\nu\|_{\mathrm{op}})e^{(\kappa-\|D\nu\|_{\mathrm{op}})t}\mathbb{E}[|Z_t|]dt \\ &\leq \frac{1}{2}(C_0((\mathrm{Ric}+\mathrm{Hess}^{\phi})^{\#}) + C_0(D\nu))e^{-(\kappa+\|D\nu\|_{\mathrm{op}})t}\rho(x,y)dt. \end{aligned}$$

Integrating both sides with the initial condition that $Z_0 = \Pi_{x,v}^{-1}W_0(u) - W_0'(w) = w - w = 0$ yields

$$\mathbb{E}[|Z_t|] \leq \frac{1}{2}(C_0((\mathrm{Ric}+\mathrm{Hess}^{\phi})^{\#}) + C_0(D\nu))\rho(x,y)\frac{e^{(\kappa-\|D\nu\|_{\mathrm{op}})t} - e^{-2\kappa t}}{\kappa - \|D\nu\|_{\mathrm{op}}}.$$

Hence, recalling that $Z_t = \Pi_{X_t,V_t}W_t - W_t'$ and the original objective to bound the integral (5.32), by assuming that $\kappa > \|D\nu\|_{\mathrm{op}}$, we find that

$$\begin{aligned} \int_0^\infty \mathbb{E}[|Z_t|]dt &\leq \frac{1}{2}(C_0((\mathrm{Ric}+\mathrm{Hess}^{\phi})^{\#}) + C_0(D\nu))\rho(x,y)\left(\frac{1}{\kappa^2 - \|D\nu\|_{\mathrm{op}}^2} - \frac{1}{2\kappa(\kappa + \|D\nu\|_{\mathrm{op}})}\right) \\ &\leq \frac{1}{4}(C_0((\mathrm{Ric}+\mathrm{Hess}^{\phi})^{\#}) + C_0(D\nu))\rho(x,y)\frac{1}{\kappa(\kappa - \|D\nu\|_{\mathrm{op}})}. \end{aligned}$$

Therefore,

$$\begin{aligned} &\int_0^\infty \mathbb{E}[|dh(Y_t)(\Pi_{X_t,V_t}^{-1}W_t) - dh(Y_t)(W_t')|]dt \\ &\qquad\leq \frac{C_0(h)}{4\kappa(\kappa - \|D\nu\|_{\mathrm{op}})}(C_0((\mathrm{Ric}+\mathrm{Hess}^{\phi})^{\#}) + C_0(D\nu))\rho(x,y). \quad (5.37) \end{aligned}$$

On combining (5.31) and (5.32), we obtain an upper bound on the Lipschitz constant of $\|D^2 f_h\|$:

$$\begin{aligned}|(df_h(x) - \Pi_{x,V} df_h(y))(u)| \leq & \int_0^\infty \mathbb{E}[|(dh(X_t) - \Pi_{X_t,V_t} dh(Y_t))(W_t)|]dt \\ & + \int_0^\infty \mathbb{E}[|dh(Y_t)(\Pi_{X_t,V_t}^{-1} W_t) - dh(Y_t)(W_t')|]dt, \\ \leq & \frac{C_1(h)}{2\kappa}\rho(x,y) \\ & + \frac{C_0(h)}{4\kappa(\kappa - \|D\nu\|_{\text{op}})}(C_0((\text{Ric} + \text{Hess}^\phi)^\#) + C_0(D\nu))\rho(x,y).\end{aligned}$$

We have established the following result:

Theorem 5.5.5. *Assume that* $\kappa > \|D\nu\|_{\text{op}}$ *and* $C_0((\text{Ric}+\text{Hess}^\phi)^\#)+C_0(D\nu) < \infty$. *Then*

$$C_1(f_h) \leq \frac{C_1(h)}{2\kappa} + \frac{C_0(h)}{4\kappa(\kappa - \|D\nu\|_{\text{op}})}(C_0((\text{Ric} + \text{Hess}^\phi)^\#) + C_0(D\nu)).$$

Example 5.5.6. For the very special case that we are working on a half plane, $M = \mathbb{R}^n_+ = \mathbb{R}^{n-1} \times \mathbb{R}_+$, many terms previously accounted for vanish. Most importantly, the compact manifold assumption can be dropped. We are now working on a Ricci flat space, Ric $= 0$, and moreover, $\nu = (0, 0, ..., 0, 1)$ remains constant along the boundary, $\partial M = \mathbb{R}^{n-1} \times \{0\}$ giving $D\nu = 0$ everywhere on ∂M. Since g is just the flat metric, the musical isomorphism of the Ricci-Bakry-Emery tensor — now just the Hessian — is invariant; $(\text{Hess}^\phi)^\# = \text{Hess}^\phi$. Consequently, the damped stochastic parallel translation is written

$$DW_t = -\frac{1}{2}\text{Hess}^\phi(W_t)dt, \quad W_0 = \text{Id}. \tag{5.38}$$

A keen eye will observe that this is just the stochastic damped parallel translation of an unreflected diffusion on $\mathbb{R}^n$. With the local time term eliminated, bounds on $\mathbb{E}[|W_t|^q]$ and $C_1(f_h)$ are significantly less complicated to calculate. Moreover, with no random part present, the stochastic covariant differential equation that governs the dynamics of the damped stochastic parallel translation is deterministic. In other words, we have the ordinary differential equation

$$\frac{dW_t}{dt} = -\frac{1}{2}\mathrm{Hess}^{\phi}(W_t).$$

Corollary 5.5.7. *Let $M = \mathbb{R}^{n-1} \times \mathbb{R}_+$ and f_h the solution to the Stein equation* (5.17)*. Assume that ϕ is 2κ-strongly log-concave and $C_2(\phi) < \infty$, then*

$$C_1(f_h) \leq \frac{C_1(h)}{2\kappa} + \frac{C_0(h)}{4\kappa^2}C_2(\phi).$$

We note here that, where we originally had $C_0((\mathrm{Hess}^{\phi})^{\#}) = C_0(\mathrm{Hess}^{\phi})$, we may simplify by seeing that

$$C_0(\mathrm{Hess}^{\phi}) = \sup_{x,y,|u|=1}\left|\frac{\mathrm{Hess}^{\phi}(x)(u) - \mathrm{Hess}^{\phi}(y)(u)}{\rho(x,y)}\right| = C_2(\phi).$$

5.6 Bounding the Wasserstein Metric

We are now at the point where we are able to bound the metric. We define a new set $\mathcal{W} = \{h \in C^1(M) : C_0(h) \leq 1, C_1(h) \leq 1\}$ as a subset of the Wasserstein set $W = \{h \in C(M) : C_0(h) \leq 1\}$. The additional smoothing requirement is to ensure that $C_1(f_h)$ is bounded. We can now confirm that for functions in $\mathcal{W}$, the Stein operator is bounded above, and moreover, the pseudo-metric associated with $\mathcal{W}$ is also finite. Recall the bounds on f_h:

$$\begin{aligned}\sup_{h\in\mathcal{W}}|\mathcal{A}f_h| &= \sup_{h\in\mathcal{W}}\left|\Delta_M f_h - \frac{1}{2}g(\nabla\phi, \nabla f_h)\right| \\ &\leq \sup_{h\in\mathcal{W}}|\Delta_M f_h| + \frac{1}{2}\sqrt{|\nabla\phi|}\sqrt{|\nabla f_h|}, \\ &\leq \sup_{h\in\mathcal{W}} C_1(f_h) + \frac{\sqrt{C_0(\phi)}}{2}\sqrt{C_0(f_h)} \\ &\leq \frac{1}{4\kappa(\kappa - \|D\nu\|_{\mathrm{op}})}(C_0((\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#}) + C_0(D\nu)) \\ &\quad + \frac{1}{2\kappa} + \frac{\sqrt{C_0(\phi)}}{2\sqrt{\kappa}} < \infty,\end{aligned}$$

so long as $\nabla\phi$ has finite norm. This, however, is already satisfied by the Lipschitz continuity assumption of the drift function.

Though we will have an upper bound for the metric $d_{\mathcal{W}}$, this is not ideal since we can not infer weak convergence (convergence in distribution) of two probability measures if $d_{\mathcal{W}} \to 0$. Instead it would be more beneficial if we could somehow forget about $C_1(f_h)$ so that we could instead bound d_W. This can be in fact achieved by comparing two Stein operators in order to admit an upper bound on the Wasserstein metric.

The idea behind this method is as follows: Suppose we have two measures $X \sim \mu$ and $Y \sim \lambda$ on M with respective densities $d\mu \propto e^{-\phi}d\text{vol}$ and $d\lambda \propto e^{-\psi}d\text{vol}$. We assume that both ϕ and ψ satisfy the sufficient condition (5.6) for the same value κ. Then for each measure we can follow through with the theory presented in this chapter to obtain a Stein operator, Stein equation and solution to said equation; X and Y will have respective Stein operators

$$\begin{aligned}\mathcal{A}_1 f &= \frac{1}{2}\Delta_M f - \frac{1}{2}g(\nabla\phi, \nabla f),\\ \mathcal{A}_2 f &= \frac{1}{2}\Delta_M f - \frac{1}{2}g(\nabla\psi, \nabla f).\end{aligned}$$

Now we draw attention to the fact that both $\mathcal{A}_1$ and $\mathcal{A}_2$ both contain the Laplace-Beltrami operator. This means that we can subtract one from the other to give us a new first order operator

$$Lf := (\mathcal{A}_2 - \mathcal{A}_1)f = \frac{1}{2}g(\nabla(\phi - \psi), \nabla f).$$

We now define f_h to be the solution to the Stein equation $\mathcal{A}_2 f_h = h - \mathbb{E}_\lambda[h(Y)]$. Then since f_h and its derivatives are bounded above, $\mathbb{E}_\mu[|\mathcal{A}_1 f_h|] < \infty$ and, more importantly, $\mathbb{E}_\mu[\mathcal{A}_1 f_h] = 0$. As a result of this, we have the following from the Stein equation:

$$\mathbb{E}_\mu[h(X)] - \mathbb{E}_\lambda[h(Y)] = \mathbb{E}_\mu[\mathcal{A}_2 f_h]$$

$$
\begin{aligned}
&= \mathbb{E}_\mu[\mathcal{A}_2 f_h] - \mathbb{E}_\mu[\mathcal{A}_1 f_h] \\
&= \mathbb{E}_\mu[L f_h]. \qquad (5.39)
\end{aligned}
$$

Whence, our problem with bounding the metric has reduced to bounding a first order operator as opposed to a second order operator. Consequently, we do not require a bound on the second derivative $C_1(f_h)$ and can therefore use the Wasserstein set W as opposed to its smoother subset $\mathcal{W}$. We now apply absolute value to both sides and take the supremum in W in Equation (5.39) to obtain the Wasserstein metric

$$d_W(X,Y) = \sup_{h \in W} |\mathbb{E}_\mu[L f_h]|.$$

We may now apply Lemma 5.4.1 in order to bound this from above:

$$d_W(X,Y) \leq \frac{1}{2\kappa} \mathbb{E}_\mu[|\nabla(\phi - \psi)(X)|].$$

Theorem 5.6.1. *Suppose two probability distributions X and Y on M with density functions proportional to $e^{-\phi}$ and $e^{-\psi}$ respectively satisfy the sufficient condition* (3.4) *for the same value κ. Then we have the following upper bound on the Wasserstein metric:*

$$d_W(X,Y) \leq \frac{1}{2\kappa} \mathbb{E}[|\nabla(\phi - \psi)(X)|].$$

In the following example we demonstrate a use for this upper bound for hemispherical data.

Example 5.6.2. In order to work with distributions on the hemisphere, we must make a modification to our manifold. Let M be a spherical cap of $N = \mathbb{S}^n$ with the restriction that $\rho \in [0, \pi/2 - \epsilon]$ for some $\epsilon > 0$. The reason that we modify the hemisphere is so that no point in M is antipodal to any other point. This means that no conjugate points exist if we are to inherit the geometry from $\mathbb{S}^n$. In practice, one should choose ϵ to be small as possible so that M is sufficiently close to the hemisphere in order to minimize the loss of information from this approximation.

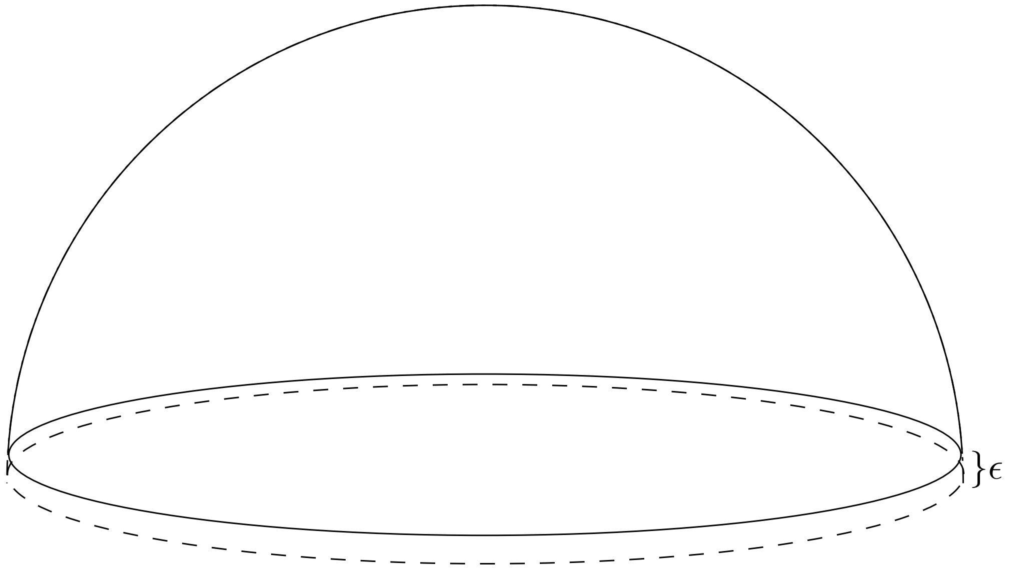

Figure 5.4: Diagram of the small cap in which we take an ϵ length off of the hemisphere.

Suppose we wish to compare a hemisphere von-Mises Fisher distribution to the uniform probability measure on M. Let $\phi = -\lambda \cos\rho$ and $\psi = 0$ be the log densities for the hemisphere VMF and uniform measure respectively. In Section 4.1 we found that for the VMF, $\text{Hess}^\phi = \lambda \cos\rho g$. Since ρ is restricted to be in $[0, \pi/2 - \epsilon]$, the lower bound on the Hessian is 0: $\text{Hess}^\phi > 0$. In contrast the VMF on $\mathbb{S}^n$, the hemisphere VMF distribution satisfies the sufficient condition (3.4) for all $\lambda > 0$. Moreover, this also satisfies the assumptions of Theorem 5.5.5 since $\nu = \partial_\rho$, $|\nu| = 1$, $\|D\nu\|_{\text{op}} = 1$ and so one must have the sufficient condition satisfied for $\kappa > 1$. This, however, is automatically satisfied since $\text{Ric} + \text{Hess}^\phi > 2g$. The sufficient condition for the uniform measure on M is also satisfied since $\text{Hess}^\psi = 0$.

For the normalizing constant of the uniform measure, we must calculate the integral $\int_0^{\pi/2-\epsilon} \sin^{n-1}\rho \, d\rho$. We find that

$$\int_0^{\pi/2-\epsilon} \sin^n \rho \, d\rho = \frac{\sqrt{\pi}}{2} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n+2}{2}\right)} - {}_2F_1\left(\frac{1}{2}, \frac{1-n}{2}, \frac{3}{2}, \sin^2\epsilon\right) \sin\epsilon, \tag{5.40}$$

where ${}_2F_1$ is the hypergeometric function defined by

$$ {}_2F_1(a,b,c,z) = \sum_{n=0}^{\infty} \frac{(a)_n(b)_n}{(c)_n}\frac{z^n}{n!} $$

where $(a)_0 = 1$ and $(a)_n = a(a+1)...(a+n-1)$ for $n > 0$.

We now apply Theorem 5.6.1 to obtain an upper bound on the Wasserstein metric,

$$ d_W(X,Y) \leq \frac{1}{2\kappa}\lambda \frac{\frac{\sqrt{\pi}}{2}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n+2}{2}\right)} - {}_2F_1\left(\frac{1}{2},\frac{1-n}{2},\frac{3}{2},\sin^2\epsilon\right)\sin\epsilon}{\frac{\sqrt{\pi}}{2}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)} - {}_2F_1\left(\frac{1}{2},\frac{2-n}{2},\frac{3}{2},\sin^2\epsilon\right)\sin\epsilon} $$

for some $\kappa > 0$.

We expect that

$$ \lim_{\epsilon\to 0}\int_0^{\pi/2-\epsilon} \sin^n\rho\, d\rho = \frac{\sqrt{\pi}}{2}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n+2}{2}\right)} $$

since sine is symmetric about $\pi/2$ and so we can clearly halve the result in Equation (4.8). To show this holds approximately, we show that the second expression on the right hand side of (5.40) is $O(\epsilon)$. It follows from the definition of ${}_2F_1$ that the derivative of ${}_2F_1$ satisfies

$$ \frac{\partial}{\partial x}\,{}_2F_1(a,b,a+1,x) = \frac{a((1-x)^{-b} - {}_2F_1(a,b,a+1,x))}{x}. $$

Denoting ${}_2F_1\left(\frac{1}{2},\frac{1-n}{2},\frac{3}{2},\sin^2\epsilon\right)$ simply as ${}_2F_1(\sin^2(\epsilon))$, we differentiate the right most expression in (5.40) to give

$$ \begin{aligned} \frac{\partial}{\partial\epsilon}\,{}_2F_1(\sin^2\epsilon)\sin\epsilon &= {}_2F_1(\sin^2\epsilon)\cos\epsilon \\ &\quad + \frac{1}{2}((1-\sin^2\epsilon)^{(n-1)/2} - {}_2F_1(\sin^2\epsilon))\sin\epsilon\frac{\sin 2\epsilon}{\sin^2\epsilon} \\ &= {}_2F_1(\sin^2\epsilon)\cos\epsilon + (\cos^{n-1}\epsilon - {}_2F_1(\sin^2\epsilon))\cos\epsilon \end{aligned} $$

$$= \cos^n \epsilon = 1 - O(\epsilon^{2n})$$

Hence, the derivative is approximately 1 for small ϵ. Moreover, ${}_2F_1(\sin^2 \epsilon)) \sin \epsilon = O(\epsilon)$ for small ϵ. Therefore, assuming continuity of the upper bound in ϵ, taking the limit as $\epsilon \to 0$ we conjecture an upper bound on the Wasserstein metric between the uniform measure and VMF distribution on the hemisphere:

$$d_W(X, Y) \leq \frac{\lambda}{2\kappa} \frac{\Gamma\left(\frac{n+1}{2}\right)^2}{\Gamma\left(\frac{n+2}{2}\right)\Gamma\left(\frac{n}{2}\right)}$$

for some $\kappa > 0$.

5.7 Conclusion

In this chapter, we have extended the Stein's method framework in [LLBF22] to the allow for distributional comparison on manifolds with a boundary.

The idea was to follow the steps taken in [LLBF22] where we include a local time term into our main SDE (5.7). The inclusion of the local time term led to a handful of complications that had to be resolved.

We found that the infinitesimal generator of the SDE with the normal reflection term (5.7) had the same infinitesimal generator as the unreflected process (5.4) under the restriction that the derivatives of functions on the boundary must vanish in the normal direction. This is akin to a Neumann boundary condition on the function space.

The coupled set of diffusions (X_t, Y_t) was formulated in such a way that the distance $\rho(X_t, Y_t)$ decreased exponentially with time. We then used this idea of coupling in the same way as in [LLBF22] to construct and solve the Stein equation.

The process of solving the Stein equation and bounding its first derivative $C_0(f_h)$ ran along the same lines as the boundaryless case. However when it came to the second derivative $C_1(f_h)$, differences with the inclusion of the normal reflection became more pronounced. The damped stochastic parallel translation now also had a reflection term present and so it was necessary to prove a Bismut—Elworth—Li formula for taking the exterior derivative inside the semigroup.

With this achieved, we moved on to bounding $C_1(f_h)$, however troubles with the local time term made it unavoidable to use the local time approximation in [AL17]. This aided in the formulation of 5.5.5 by making the dL_t term into a dt term that had infinite push when approaching the boundary.

On comparing the bound on the second derivative $C_1(f_h)$ from the boundaryless case in [LLBF22] to the boundary case in Theorem 5.5.5 we find an improvement: We have in the boundaryless case,

$$C_1(f_h) \leq \frac{C_1(h)}{2\kappa} + \frac{C_0(h)}{2\kappa^2} C_0((\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#}),$$

and in the boundary case,

$$C_1(f_h) \leq \frac{C_1(h)}{2\kappa} + \frac{C_0(h)}{4\kappa(\kappa - \|D\nu\|_{\mathrm{op}})} (C_0((\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#}) + C_0(D\nu)).$$

On setting $\partial M = 0$, $\|D\nu\|_{\mathrm{op}} = 0$, we have

$$C_1(f_h) \leq \frac{C_1(h)}{2\kappa} + \frac{C_0(h)}{4\kappa^2} C_0((\mathrm{Ric} + \mathrm{Hess}^{\phi})^{\#}),$$

an improvement in the second term by 1/2. Moreover, this also improves the second derivative result in [MG16] when setting $M = \mathbb{R}^n$, $\mathrm{Ric} = 0$.

Finally, it is possible to imply weak convergence without using the Wasserstein metric on M. By [BOPG18, Corollary 1], taking $w_n^{(n)} = 1$ and the rest of the weights to be 0, if we can show that the kernel Stein discrepancy $\sup_{f \in \mathcal{H}} \mathbb{E}_{P_n}[|\mathcal{A}f|] \to 0$ as $n \to \infty$ where $\mathcal{H}$ is the unit ball of a reproducing kernel Hilbert space, then we can imply weak convergence. However, since we haven't shown that the expectation of the operator of a kernel converges to 0 in this work, this result could be used instead if stronger assumptions on the smoothness of our test functions were required.

Chapter 6

Conclusion and Future Work

6.1 Summary

This thesis has been a thorough showcase for Stein's method on manifold. We have been rather comprehensive in our discussion, applying both the density and diffusion approaches to different problems.

In Chapter 3, we proposed a density approach for comparing distributions on $\mathbb{S}^1$. This was motivated by the fact that a wrapping was needed when defining $\mathbb{S}^1$ in terms of an interval. We introduced the Stein operator and its inverse, necessary for assembling the Stein equation. We found it was necessary to redefine the Stein kernel due to the fact that $\mathbb{E}[X]$ is not uniquely defined on $\mathbb{S}^1$. We did this by introducing the circular Stein kernel τ^c. With these tools in hand, Theorem 3.3.2 was formulated, giving us an upper and lower bound on the Wasserstein metric for distributions on $\mathbb{S}^1$. Analytic examples of distributional comparisons were given in Section 3.3.2, two Bayesian model comparison, and a comparison between the wrapped normal and wrapped Cauchy distributions. It was noted that the formula for the bound on the Wasserstein metric does not provide us with robust, sharp analytic bound for all distributions, for example between the von-Mises and the Bingham distributions. Instead, we relied upon numerical integration schemes to approximate the upper bound of the Wasserstein metric. This was done for the wrapped normal and von-Mises distribution, and the resulting approximation was compared to an asymptotic result of Kent.

In Chapter 4, we used the upper bound on the Wasserstein metric from [LLBF22], Theorem 4.0.1, to compare numerous distributions on a variety of manifolds. We looked at three comparisons on $\mathbb{S}^n$: between the uniform and von-Mises; between the von-Mises and von-Mises; and between the uniform and Fisher–Watson distributions. Particularly, we expressed the comparison between two von-Mises Fisher distributions $\mathrm{VMF}(\mu_1, \lambda_1)$ and $\mathrm{VMF}(\mu_2, \lambda_2)$ in terms of an expectation. Through this, we could then more easily look at the case when $\mu_1 = \mu_2$. In Section 4.2, our objective was to compare the heat kernel of $\mathbb{H}^3$ and the Riemannian-Gaussian distribution. We were able to obtain a bound on the Wasserstein metric for $t < \frac{1}{4}$, and relate this to Varadhan's asymptotic relation. Effectively, this is a special case where we have proved Varadhan's asymptotic relation for finite t. We next moved on to $\mathrm{SO}(n)$ in Section 4.3, comparing the uniform measure and the matrix von-Mises distributions and two matrix von-Mises distributions. We were able to obtain a bound in terms of an expectation in both cases, but for the uniform it was possible to numerically approximate this by sampling from the Haar measure on $\mathrm{SO}(n)$. We finished the chapter with $\mathcal{P}_n$ which took some more geometric work in order to prove that the Bakry-Èmery-Ricci criterion was satisfied for the Riemannian-Gaussian distribution. We then finished the chapter by comparing two Riemannian-Gaussian distributions with different distributional parameters.

The last chapter, Chapter 5, provided a framework for Stein's method on manifolds with a boundary. The work of [LLBF22] assumed that $\partial M = \emptyset$, and so this was an obvious direction to head towards next. We did this by extending the framework of [LLBF22], adding in a reflecting local time term to the stochastic differential equation. We noted that adding a local time correction in the SDE of the diffusion was required to account for the boundary. We began in Section 5.2 by trying to determine what the invariant distribution is and working out what the infinitesimal generator of said diffusion was. We found that to formulate the infinitesimal generator was only possible in the case where we restricted the support of the generator to be $\{f \in C^2(M) : df(\nu) = 0\}$. In this case, we obtained the correct invariant distribution and were able to use the infinitesimal generator as a Stein operator. We then moved on to the construction of the coupled diffusions in Section 5.3. We showed that even with the inclusion of the local time term, we still retained the coupling result from the boundary-less case. With the Stein operator and coupling in hand, we formulated the Stein equation and found its solution f_h in Section 5.4. We were also able to bound $C_0(f_h)$ and $\|f_h\|_\infty$. To bound $C_1(f_h)$ in Section 5.5.3 required a Bismut–Elworthy–Li formula to exchange the exterior derivative and expectation. We therefore needed the Weitzenböck formula and the notion of damped stochastic parallel displacement to prove this. With a Bismut–Elworthy–Li formula in hand, it was then possible to bound $C_1(f_h)$, and then in Section 5.6, the Wasserstein metric. We finished with an example in which we compared the uniform measure and the von-Mises Fisher distribution when restricted to the small cap on $\mathbb{S}^n$.

6.2 Future Work

Since Stein's method has been shown to be increasingly more applicable in statistics, it is obvious that a possible future research topic is to look at the statistical applications of the framework for manifolds with boundary. Barp et. al. [BOPG18] had previously done this for numerical approximation of integrals, but their approaches and assumptions diverge from ours. With their assumption that the density p vanishes on the boundary, the space of test functions is easy to work with. With our new approach, we have instead restricted the space of test functions, therefore we may not be able to infer distributional convergence from the kernel Stein discrepancy (mentioned in the Introduction). Exploring whether we may still be able to infer this is an interesting question. If so, a wider range of probability measures may be applicable for comparison.

The positive orthant of $\mathbb{S}^n$ a is space of interest for statisticians and it may be possible to use this manifold with boundary under our framework. Despite the boundary being non-differentiable at a finite number of points, since ∂M is codimension 1 with M, a diffusion will not hit these sharp points with probability 1. The consequence of this observation is that it might be possible to use our framework for the positive orthant, or even a general manifold with differentiable boundary everywhere apart from a measure-zero set of points.

Higher order bounds on the solution to the Stein equation rapidly increase the difficulty of calculation. But for the half plane case, many calculations simplify with the absence of the local time term. Generalizations and improvements to the results of [MG16] are possible by utilising the damped stochastic parallel translation. For $C_2(f_h)$, doubly damped stochastic parallel translations [Li16] are required to tackle this. These are derivatives of the original damped stochastic parallel translation in a different direction. Bismut–Elworthy–Li formulae exist [Tho20] for the second derivative, and with $\mathrm{Ric} = 0$, have nice forms.

Properties of the kernel Stein discrepancy have allowed the weakening of assumptions of probability measures and classes of test functions whilst also bounding integral probability metrics like the Wasserstein metric, [BOPG18, GDVM19]. It is clear the kernel Stein discrepancy is an exceptionally strong tool that warrants further research. Particularly, it would be of interest to look at the diffusion Stein kernel and see if it is possible to bound the Wasserstein metric using it, with the hope that assumptions like the Bakry-Èmery-Ricci criterion can be relaxed.

Bibliography

[AL17] Marc Arnaudon and Xue-Mei Li. Reflected Brownian Motion: Selection, Approximation and Linearization. *Electronic Journal of Probability*, 22:1–55, 2017.

[And03] Attila Andai. *Information Geometry in Quantum Mechanics.* PhD thesis, Budapest University of Technology and Economics, 2003.

[And11] Sebastian Andres. Pathwise Differentiability for SDEs in a Smooth Domain with Reflection. *Electronic Journal of Probability*, 16:845–879, 2011.

[Aub12] Thierry Aubin. *Nonlinear Analysis on Manifolds. Monge-Ampere Equations*, volume 252. Springer Science & Business Media, 2012.

[Bak86] Dominique Bakry. Un Critère de Non-explosion Pour Certaines Diffusions sur une Variété Riemannienne Complète. *Comptes rendus de l'Académie des sciences. Série 1, Mathématique*, 303(1):23–26, 1986.

[Bar90] Andrew D Barbour. Stein's Method for Diffusion Approximations. *Probability theory and related fields*, 84(3):297–322, 1990.

[BOPG18] Alessandro Barp, Chris Oates, Emilio Porcu, and Mark Girolami. A Riemannian-Stein Kernel Method. *arXiv preprint arXiv:1810.04946*, 1(5):6–9, 2018.

[BP99] Timothy C Brown and Michael J Phillips. Negative Binomial Approximation with Stein's Method. *Methodology and computing in applied probability*, 1(4):407–421, 1999.

[CE08] Jeff Cheeger and David G Ebin. *Comparison Theorems in Riemannian Geometry*, volume 365. American Mathematical Society, 2008.

[CFR11] Sourav Chatterjee, Jason Fulman, and Adrian Röllin. Exponential Approximation by Stein's Method and Spectral Graph Theory. *ALEA Latin American Journal of Probability and Mathematical Statistics*, 8(1):197–223, 2011.

[CGS10] Louis HY Chen, Larry Goldstein, and Qi-Man Shao. *Normal Approximation by Stein's Method*. Springer Science & Business Media, 2010.

[Che75] Louis HY Chen. Poisson Approximation for Dependent Trials. *The Annals of Probability*, 3(3):534–545, 1975.

[CSG16] Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A Kernel Test of Goodness of Fit. In *International conference on machine learning*, pages 2606–2615. Proceedings of Machine Learning Research, 2016.

[DF87] Persi Diaconis and David Freedman. A Dozen de Finetti-style Results in Search of a Theory. In *Annales de l'IHP Probabilités et statistiques*, volume 23, pages 397–423, 1987.

[Döb15] Christian Döbler. Stein's Method of Exchangeable Pairs for the Beta Distribution and Generalizations. *Electronic Journal of Probability*, 20:1–34, 2015.

[DP18] Alberto Dolcetti and Donato Pertici. Differential Properties of Spaces of Symmetric Real Matrices. *arXiv preprint arXiv:1807.01113*, 2018.

[DW99] Paul Damien and Stephen Walker. A Full Bayesian Analysis of Circular Data Using the von–Mises Distribution. *Canadian Journal of Statistics*, 27(2):291–298, 1999.

[EK09] Stewart N Ethier and Thomas G Kurtz. *Markov Processes: Characterization and Convergence*, volume 282. John Wiley & Sons, 2009.

[Fre16] Mark Iosifovich Freidlin. *Functional Integration and Partial Differential Equations.(AM-109), Volume 109.* Princeton university press, 2016.

[GDVM19] Jackson Gorham, Andrew B Duncan, Sebastian J Vollmer, and Lester Mackey. Measuring Sample Quality with Diffusions. *The Annals of Applied Probability*, 29(5):2884–2928, 2019.

[GH80] Donald Geman and Joseph Horowitz. Occupation Densities. *The Annals of Probability*, pages 1–67, 1980.

[GHL90] Sylvestre Gallot, Dominique Hulin, and Jacques Lafontaine. *Riemannian Geometry*, volume 2. Springer, 1990.

[GL88] Peter Guttorp and Richard A Lockhart. Finding the Location of a Signal: A Bayesian Analysis. *Journal of the American Statistical Association*, 83(402):322–330, 1988.

[GL18] Fatemeh Ghaderinezhad and Christophe Ley. A General Measure of the Impact of Priors in Bayesian Statistics via Stein's Method. *arXiv preprint arXiv:1803.00098*, 2018.

[GM15] Jackson Gorham and Lester Mackey. Measuring Sample Quality with Stein's Method. *Advances in Neural Information Processing Systems*, 28, 2015.

[GN98] Alexander Grigor'yan and Masakazu Noguchi. The Heat Kernel on Hyperbolic Space. *Bulletin of the London Mathematical Society*, 30(6):643–650, 1998.

[GW06] Robert Everist Greene and Hung-Hsi Wu. *Function Theory on Manifolds Which Possess a Pole*, volume 699. Springer, 2006.

[Hsu02] Elton P Hsu. *Stochastic Analysis on Manifolds*. Number 38. American Mathematical Society, 2002.

[IW14] Nobuyuki Ikeda and Shinzo Watanabe. *Stochastic Differential Equations and Diffusion Processes*. Elsevier, 2014.

[Jos08] Jürgen Jost. *Riemannian Geometry and Geometric Analysis*, volume 42005. Springer, 2008.

[Ken75] John Kent. Discussion of Professor Mardia's Paper. *Journal of the Royal Statistical Society: Series B*, 37:371–393, 1975.

[Ken86] Wilfrid S Kendall. Stochastic Differential geometry, a Coupling Property, and Harmonic Maps. *Journal of the London Mathematical Society*, 2(3):554–566, 1986.

[KR14] Weining Kang and Kavita Ramanan. Characterization of Stationary Distributions of Reflected Diffusions. *The Annals of Applied Probability*, 24(4):1329–1374, 2014.

[Kro79] Stephen Kronwith. Convex Manifolds of Nonnegative Curvature. *Journal of Differential Geometry*, 14(4):621–628, 1979.

[Lee18] John M Lee. *Introduction to Riemannian Manifolds*. Springer, 2018.

[Li92] Xue-Mei Li. *Stochastic Flow on Noncompact Manifolds*. PhD thesis, University of Warwick, 1992.

[Li16] Xue-Mei Li. Doubly Damped Stochastic Parallel Translations and Hessian Formulas. In *International Conference on Stochastic Partial Differential Equations and Related Fields*, pages 345–357. Springer, 2016.

[LLBF22] Huiling Le, Alexander Lewis, Karthik Bharath, and Christopher Fallaize. A Diffusion Approach to Stein's Method on Riemannian Manifolds. *arXiv preprint arXiv:2003.11497*, 2022.

[LLJ16] Qiang Liu, Jason Lee, and Michael Jordan. A Kernelized Stein Discrepancy for Goodness-of-fit Tests. In *International conference on machine learning*, pages 276–284. Proceedings of Machine Learning Research, 2016.

[Loh92] Wei-Liem Loh. Stein's Method and Multinomial Approximation. *The Annals of Applied Probability*, 2(3):536–554, 1992.

[LRS17a] Christophe Ley, Gesine Reinert, and Yvik Swan. Distances Between Nested Densities and a Measure of the Impact of the Prior in Bayesian Statistics. *The Annals of Applied Probability*, 27(1):216–241, 2017.

[LRS17b] Christophe Ley, Gesine Reinert, and Yvik Swan. Stein's Method for Comparison of Univariate Distributions. *Probability Surveys*, 14:1–52, 2017.

[Luk94] Ho Ming Luk. *Stein's Method for the Gamma Distribution and Related Statistical Applications.* PhD thesis, University of Southern California, 1994.

[LW16] Qiang Liu and Dilin Wang. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. *arXiv preprint arXiv:1608.04471*, 2016.

[MEA76] Kanti V Mardia and SAM El-Atoum. Bayesian Inference for the von–Mises-Fisher Distribution. *Biometrika*, 63(1):203–206, 1976.

[Mec09] Elizabeth Meckes. On the Approximate Normality of Eigenfunctions of the Laplacian. *Transactions of the American Mathematical Society*, 361(10):5377–5399, 2009.

[MG16] Lester Mackey and Jackson Gorham. Multivariate Stein Factors for a Class of Strongly Log-Concave Distributions. *Electronic Communications in Probability*, 21, 2016.

[MJ09] Kanti V Mardia and Peter E Jupp. *Directional Statistics*, volume 494. John Wiley & Sons, 2009.

[MRS18] Guillaume Mijoule, Gesine Reinert, and Yvik Swan. Stein Operators, Kernels and Discrepancies for Multivariate Continuous Distributions. *arXiv preprint arXiv:1806.03478*, 2018.

[MSW19] Ovidiu Munteanu, Chiung-Jue Anna Sung, and Jiaping Wang. Poisson Equation on Complete Manifolds. *Advances in Mathematics*, 348:81–145, 2019.

[MZ11] Maher Moakher and Mourad Zéraï. The Riemannian Geometry of the Space of Positive-definite Matrices and its Application to the Regularization of Positive-definite Matrix-valued Data. *Journal of Mathematical Imaging and Vision*, 40(2):171–187, 2011.

[NP09] Ivan Nourdin and Giovanni Peccati. Stein's Method on Wiener Chaos. *Probability Theory and Related Fields*, 145(1-2):75–118, 2009.

[NP12] Ivan Nourdin and Giovanni Peccati. *Normal Approximations with Malliavin Calculus: From Stein's Method to Universality.* Number 192. Cambridge University Press, 2012.

[NPR10] Ivan Nourdin, Giovanni Peccati, and Anthony Réveillac. Multivariate Normal Approximation Using stein's Method and Malliavin Calculus. 46(1):45–58, 2010.

[Øks13] Bernt Øksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

[Pek96] Erol A Peköz. Stein's Method for Geometric Approximation. *Journal of Applied Probability*, 33(3):707–713, 1996.

[PR12] John Pike and Haining Ren. Stein's Method and the Laplace Distribution. *arXiv preprint arXiv:1210.5775*, 2012.

[RCC$^+$20] Marina Riabiz, Wilson Chen, Jon Cockayne, Pawel Swietach, Steven A Niederer, Lester Mackey, and Chris Oates. Optimal Thinning of MCMC Output. *arXiv preprint arXiv:2005.03952*, 2020.

[Ros11] Nathan Ross. Fundamentals of Stein's Method. *Probability Surveys*, 8:210–293, 2011.

[RW94] LCG Rogers and David Williams. *Diffusions, Markov Processes and Martingales, Volume 1: Foundations*. Chichester: Wiley, 1994.

[RW00] LCG Rogers and David Williams. *Diffusions, Markov Processes and Martingales: Volume 2, Itô Calculus*, volume 2. Cambridge university press, 2000.

[SBBM17] Salem Said, Lionel Bombrun, Yannick Berthoumieu, and Jonathan H Manton. Riemannian Gaussian Distributions on the Space of Symmetric Positive Definite Matrices. *IEEE Transactions on Information Theory*, 63(4):2153–2170, 2017.

[Ste72] Charles Stein. A Bound for the Error in the Normal Approximation to the Distribution of a Sum of Dependent Random Variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California, 1972.

[Ste86] Charles Stein. Approximate Computation of Expectations. *Lecture Notes-Monograph Series*, 7:i–164, 1986.

[Tho20] James Thompson. Approximation of Riemannian Measures by Stein's Method. *arXiv preprint arXiv:2001.09910*, 2020.

[Wan94] Feng-Yu Wang. Application of Coupling Methods to the Neumann Eigenvalue Problem. *Probability Theory and Related Fields*, 98(3):299–306, 1994.

[Wan14] Feng-Yu Wang. *Analysis for Diffusion Processes on Riemannian Manifolds*, volume 18. World Scientific, 2014.

[Xia98] Yimin Xiao. Local Times and Related Properties of Multidimensional Iterated Brownian Motion. *Journal of Theoretical Probability*, 11(2):383–408, 1998.

Appendix A

Variations and Jacobi Fields

Suppose we have a geodesic $\gamma : [0,1] \to M$ connecting two points x and y on M, then we can construct a two-parameter family of curves which connect these two points. If we define $f : [0,1] \times (-\epsilon, \epsilon) \to M$ for $\epsilon > 0$ such that $f(0,s) = x$, $f(1,s) = y$, $\forall s \in (-\epsilon, \epsilon)$ and $f(t,0) = \gamma(t)$, $\forall t \in [0,1]$. These are called variations of the geodesic.

On some manifolds these variations manifest as geodesics. If we take $M = \mathbb{S}^n$, $x \in \mathbb{S}^n$ and let y be anti-podal to x. Then y is the conjugate point of x and there are infinitely many geodesics connecting the two. If we are to fix one γ and define the variations as γ rotated by the angle ϕ then these curves are clearly also geodesics.

Suppose we now have two orthogonal vector fields at $f(t,s)$ defined by $T = \partial_t f(t,s)$ and $V = \partial_s f(t,s)$. By the torsion-free property of the Levi-Civita connection, $D_T V = D_V T$ ($[T,V] = 0$ due to orthogonality) and so we may write that $D_T D_T V = D_T D_V T$. Evaluating this at $s = 0$, we can further simplify by noting that $T|_{s=0} = \dot{\gamma}$ and hence

$$D_{\dot{\gamma}} D_{\dot{\gamma}} V = D_{\dot{\gamma}} D_V \dot{\gamma} = D_{\dot{\gamma}} D_V \dot{\gamma} - D_V D_{\dot{\gamma}} \dot{\gamma} = R(\dot{\gamma}, V)\dot{\gamma}.$$

By labelling $D_{\dot{\gamma}} = D_t$ we obtain the Jacobi equation

$$\ddot{V} = R(\dot{\gamma}, V)\dot{\gamma}. \tag{A.1}$$

Taking a step back to where we defined f, we noted that by fixing $s = 0$ one can generate $\dot{\gamma}$ by taking derivative in t. We can also look at the case where we take the derivative in s instead. This vector field generated by this two parameter family in s is called the Jacobi field along γ, i.e.

$$J = \frac{\partial}{\partial s} f(t, s)\bigg|_{s=0}.$$

In fact, this vector field solves the Jacobi equation (A.1);

$$\begin{aligned}\ddot{J} &= D_t D_t \partial_s f|_{s=0}\\ &= D_t D_s \partial_t f|_{s=0}\\ &= R(\dot{\gamma}, J)\dot{\gamma} + D_s D_t \dot{\gamma}\\ &= R(\dot{\gamma}, J)\dot{\gamma},\end{aligned}$$

where we have labelled $D_s = D_{\partial s}$.

Jacobi fields are useful in many ways, for example Killing vector fields are Jacobi fields when restricted to geodesics. One important use of Jacobi fields is what is known as the second variation of length formula.

Theorem A.0.1 (The Second Variation Formula). *Let $L(f_s)$ denote the length of the curve $f(t, s)$ for a fixed s. Let J be the Jacobi field associated with f. Then,*

$$\frac{\partial^2 L(f_s)}{\partial s^2}\bigg|_{s=0} = \int_0^1 g(D_t J, D_t J) - g(R(J, \dot{\gamma})\dot{\gamma}, J) dt. \tag{A.2}$$

The integral on the right hand side is more commonly known as the index form and is represented as $I : \Gamma(TM) \times \Gamma(TM) \to \mathbb{R}$.

Appendix B

Lie Groups

Definition B.0.1. A Lie group G is a set G with identity element I and the following:

i) a multiplication mapping; $\mu : G \times G \to G$, $\mu(gh) = gh \in G \; \forall g, h \in G$.

ii) an inverse mapping; $\sigma : G \to G$, $\sigma(g) = g^{-1} \in G, gg^{-1} = g^{-1}g = I \; \forall g \in G$.

iii) G has the structure of a smooth manifold. Moreover, μ and σ are both smooth functions on G.

Definition B.0.2. Let M be a manifold and G be a Lie group. A left action of G on M is a map $\lambda : G \times M \to M$ such that

$$\lambda(\lambda(g, h), x) = \lambda(g, \lambda(h, x))$$

and

$$\lambda(I, x) = x$$

for all $g, h \in G$ and $x \in M$.

Definition B.0.3. The exponential map of an element S of a Lie group G is defined as

$$\exp(S) := \sum_{n=0}^{\infty} \frac{S^n}{n!}$$

where we define $S^0 := I$.

Example B.0.4. Define the matrix A as

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Then A provides a basis in the space of antisymmetric 2×2 matrices with 0 trace. To calculate its exponential, it is very useful to note that $A^2 = -I$ from which we can infer that $A^3 = -A$, $A^4 = I$ and $A^5 = A$ and so on. We then compute the following, using the fact that we can separate cases when n is odd or even;

$$\begin{aligned}
\exp(tA) &= \sum_{n=0}^{\infty} \frac{(tA)^n}{n!} \\
&= \sum_{n=0}^{\infty} \frac{(tA)^{2n+1}}{(2n+1)!} + \sum_{n=0}^{\infty} \frac{(tA)^{2n}}{(2n)!} \\
&= A \sum_{n=0}^{\infty} \frac{(-1)^n t^n}{(2n+1)!} + I \sum_{n=0}^{\infty} \frac{(-1)^n t^n}{(2n)!} \\
&= A \sin(t) + I \cos(t).
\end{aligned}$$

In matrix form this is

$$e^{tA} = \begin{pmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{pmatrix}.$$

A few more interesting properties of the matrix exponential are:

- If D is diagonal then clearly e^D will also be diagonal with exponentiated entries,
- If P is an idempotent matrix, i.e. $P^2 = P$, then $e^P = I + (e-1)P$,
- If A is diagonalizable, i.e. $A = QDQ^{-1}$ for some diagonal matrix D and the columns of Q are an orthonormal basis of eigenvectors, then $e^A = Qe^DQ^{-1}$.

Unlike the typical exponential function on $\mathbb{R}$, for matrices A and B, it is not necessarily the case that $e^A e^B = e^{A+B}$. This is generally only true when A and B commute, i.e. $AB = BA$. For general, possibly non-commuting A, B, one must make use of the famous Baker-Campbell-Hausdorff formula.

Definition B.0.5. A Lie algebra is a vector space $\mathfrak{g}$ together with a binary operation $[\cdot,\cdot] : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ called the Lie bracket satisfying the following:

i) Bilinearity; $[ax+by,z] = a[x,z]+b[y,z],\ [z,ax+by] = a[z,x]+b[z,y]$,

ii) Alternativity; $[x,x] = 0.$,

iii) Jacobi identity; $[x,[y,z]]+[z,[x,y]]+[y,[z,x]] = 0.$,

iv) Anticommutativity; $[x,y] = -[y,x]$.

For all $x,y,z \in \mathfrak{g},\ a,b \in \mathbb{R}$.

Lie algebra are directly related to Lie groups by the following characterization: If $\mathfrak{g}$ is a Lie algebra of a Lie group G, then for all $A \in \mathfrak{g}$, $e^{tA} \in G$, $\forall t \in \mathbb{R}$. This particular mapping will generate the whole group G from $\mathfrak{g}$. The elements of $\mathfrak{g}$ are sometimes called the infinitesimal generators of G. In addition to this, $\mathfrak{g}$ is isomorphic to T_IG, the tangent space at identity. In Example B.0.4, without knowing, we showed that the Lie algebra of $\mathrm{SO}(n,\mathbb{R})$ (the space of orthogonal matrices with determinant 1) is the space of skew symmetric 2×2 matrices with 0 trace. This particular characterization can be extended to $\mathrm{SO}(n,\mathbb{R})$. We abbreviate $\mathrm{SO}(n,\mathbb{R})$ to $\mathrm{SO}(n)$ as we shall not be using any field other than $\mathbb{R}$.

Example B.0.6. First, let us show that the Lie algebra of $\mathrm{SO}(n)$ generates the whole group. Denote the lie algebra of $\mathrm{SO}(n)$, the set of $n \times n$ skew symmetric matrices with 0 trace by $\mathrm{skew}(n)$. Let $A \in \mathrm{skew}(n)$, then

$$e^A\left(e^A\right)^\intercal = e^A e^{A^\intercal}.$$

Now since $A^\intercal = -A$, $[A, A^\intercal] = 0$, and hence the product of the exponentials is the exponential of the sum;

$$e^A e^{A^\intercal} = e^{A+A^\intercal} = e^0 = I.$$

Hence, $e^A \in \mathrm{SO}(n)$.

To show the converse, assume that an element of $\mathrm{SO}(n)$ can be written in the form e^{tM} for some matrix M and $t \in \mathbb{R}$. Then by definition of $\mathrm{SO}(n)$, $e^{tM}\left(e^{tM}\right)^{\intercal} = I$. Taking derivatives at 0 yields on both sides the following:

$$0 = Me^{tM}\left(e^{tM}\right)^{\intercal} + e^{tM}\left(e^{tM}\right)^{\intercal}M^{\intercal}\Bigg|_{t=0} = M + M^{\intercal},$$

which characterizes $\mathrm{skew}(n)$. Hence $M \in \mathrm{skew}(n)$. Thus, we have showed that the Lie algebra of $\mathrm{SO}(n)$ is $\mathrm{skew}(n)$.

Definition B.0.7. Let $g \in G$. The adjoint action $\mathrm{Ad}_g : \mathfrak{g} \to \mathfrak{g}$ is defined as

$$\mathrm{Ad}_g(A) = gAg^{-1}$$

It turns out that not only is the adjoint action an interesting quantity, but also its derivative $d\mathrm{Ad}_g$ is too.

By definition of the Lie algebra of G, we may write $g = e^{tC}$ for some $C \in \mathfrak{g}$, hence $\mathrm{Ad}_{e^{tC}}(A) = e^{tC}Ae^{-tC}$. Then we can easily compute the differential at identity;

$$\begin{aligned}\frac{\partial}{\partial t}\mathrm{Ad}_{e^{tC}}(A)\Bigg|_{t=0} &= \frac{\partial}{\partial t}e^{tC}Ae^{-tC}\Bigg|_{t=0} \\ &= Ce^{tC}Ae^{-tC} - e^{tC}AC^{-tC}\Bigg|_{t=0} \\ &= CA - AC = [C, A].\end{aligned}$$

Thus, we have shown that the differential at identity of the adjoint action is precisely the Lie bracket on $\mathfrak{g}$. This special quantity is widely known as the adjoint representation of $\mathfrak{g}$.

Definition B.0.8. Let $X \in \mathfrak{g}$. The adjoint representation $\mathrm{ad}_X : \mathfrak{g} \to \mathfrak{g}$ is the differential at identity of $\mathrm{Ad}_{e^{tX}}$ and is equal to the Lie bracket on $\mathfrak{g}$.

Definition B.0.9. Let $g \in G$. Then the map $L_g : G \to G$ ($R_g : G \to G$) that is defined by

$$L_g(h) = gh, \ \big(R_g(h) = hg\big), \ h \in G$$

are called left (right) translations.

Definition B.0.10. A vector field $X \in TG$ is called left invariant if for any $g \in G$

$$(DL_g)X = X \circ L_g.$$

In other words,

$$\big(D_h L_g\big)X(h) = X(gh).$$

Here, D_h is the directional derivative of L; $D_h L_g : T_h G \to T_{gh} G$.

In order to calculate left invariant vector fields (LIVF's for short), the following formula can be applied. A LIVF U of a corresponding element of the Lie algebra $A \in \mathfrak{g}$ acts on smooth functions f by

$$Uf(S) = \frac{d}{dt} f(Se^{tA})\bigg|_{t=0}$$

for $S \in G$. Note that every element in the Lie algebra is, by construction, left invariant. This can be seen by applying the above formula $f = \mathrm{Id}$ and setting $S = I$. Furthermore, using this definition, we are able to map from to any tangent space from the tangent space at identity. Let $S \in G$ and $A \in \mathfrak{g}$, then $SA \in T_S G$. In addition, the path on G, Se^{tA} characterizes a geodesic on G. Similarly to the Riemannian manifolds we have discussed previously, points on a Lie group may have cut points. That is, the exponential map is not injective everywhere. For example, take $G = \mathrm{SO}(3)$ with the bi-invariant metric $g(A, B)_I = \dfrac{1}{2}\mathrm{Tr}(B^\intercal A)$ for $A, B \in \mathfrak{so}(3)$. By the Rodrigues rotation formula, for $E \in \mathfrak{so}(3)$, we can express any element S of SO(3) via

$$S = I + \sin(t)E + (1 - \cos(t))E^2$$

for $t \in \mathbb{R}$. This formula is direct consequence of an application of the exponential map after noting that characteristic polynomial of E is $p_E(\lambda) = -\lambda^3 - \lambda$. Using this we define two geodesics on SO(3) connecting S and I; $\gamma_1(t) = Se^{\pi t E_1}$ and $\gamma_2(t) = Se^{\pi t E_2}$ $t \in [0,1]$ with the relation that $E_2 = -E_1 = E_1^\intercal$. Then via the Rodrigues formula, both $\gamma_1(1) = \gamma_2(1) = 1 + E_1^2 = S$ and the distances of the two paths $d_{\gamma_i}(I,S) = \sqrt{g(\dot{\gamma}_i(0), \dot{\gamma}_i(0))}$, $i = 1,2$ are also the same,

$$d_{\gamma_1}(I,S) = \frac{1}{\sqrt{2}}\sqrt{\mathrm{Tr}(E_1^\intercal E_1)},$$

$$d_{\gamma_2}(I,S) = \frac{1}{\sqrt{2}}\sqrt{\mathrm{Tr}(E_2^\intercal E_2)} = \frac{1}{\sqrt{2}}\sqrt{\mathrm{Tr}(E_1^\intercal E_1)}.$$

Definition B.0.11. A metric $g(\cdot,\cdot)$ on a Lie group G is called left (right) invariant iff

$$g(X,Y)_h = g\big((D_h L_g)X, (D_h L_g)Y\big)_{gh}$$

$$\big(g(X,Y)_h = g\big((D_h R_g)X, (D_h R_g)Y\big)_{hg}\big)$$

for all $h, g \in G$ and $X, Y \in T_h G$.

Definition B.0.12. A metric is called bi-invariant if it is both left and right invariant.

Lemma B.0.13. *For an Lie group G equipped with a bi-invariant metric, the following properties hold:*

i) $D_X Y = \frac{1}{2}[X,Y],\ X,Y \in \mathfrak{g}^L$,

ii) $R(U,V) = \frac{1}{4}\mathrm{ad}_{[U,V]},\ U,V \in \mathfrak{g}$, *or*
$R(U,V)W = \frac{1}{4}[[U,V],W],\ U,V,W \in \mathfrak{g}$,

iii) $K(U,V) = \frac{1}{4}g([U,V],[U,V])$ *for all orthonormal vectors* $U,V \in \mathfrak{g}$,

iv) $\mathrm{Ric}(U,V) = -\frac{1}{4}B(U,V),\ U,V \in \mathfrak{g}$
where B is the killing form defined by $B(U,V) = \mathrm{Tr}(ad_U \circ ad_V)$.

Appendix C

Orthonormal Frame Bundle

Definition C.0.1. Let E, B and F be smooth manifolds and $\pi : E \to B$ be a smooth projection mapping. The triple (π, E, B) is a fibre bundle with fibre F, base B, and total space E if:

i) the map π is surjective,

ii) there exists an open covering $\{U_i\}_{i\in I}$ of B, and diffeomorphisms

$$h_i : \pi^{-1}(U_i) \to U_i \times F$$

such that $h_i \circ \pi^{-1}(x) = \{x\} \times F$ for $x \in U_i$.

The fibre at a point x can be generated by taking the mapping $\pi^{-1}(x) = F$.

Example C.0.2. Take $B = M$, a Riemannian manifold, $E = TM$ its tangent bundle with canonical projection $\pi : (x, v) \mapsto x$. Then, as the name might suggest, (π, M, TM) is a fibre bundle with fibres at x equal to T_xM. We can take $U_i = M$ and $h(x, v) = (x, \pi^{-1}(x))$, then h satisfies part ii).

Definition C.0.3. Let M and P be manifolds, let G be a Lie group and let $\pi : P \to M$ be a smooth map. We call (P, π, M, G) a principle bundle over M with structure group G if the following three conditions are satisfied:

i) G acts freely on P on the right, i.e. there is a right action

$$P \times G \to P$$

$$(p, g) \mapsto R_g(p) = pg$$

with the property that there are no fixed points of R other than the identity on G; $R_g(p) = p$ for some $p \in P$ iff $g = I$,

ii) For $p_1, p_2 \in P$, there exists some $g \in G$ with $p_2 = R_g(p_1)$ iff $\pi(p_1) = \pi(p_2)$,

iii) The diffeomorphisms h_i in part ii) of Definition C.0.1 are isomorphisms with action of G.

To lead onto the construction of a Brownian motion process on a manifold, we introduce the orthonormal frame bundle.

Definition C.0.4. A frame at a point $x \in M$ is an $\mathbb{R}$-linear isomorphism $u_x : \mathbb{R}^d \to T_xM$. Explicitly, $u(e_i+e_j) = u(e_i)+u(e_j)$ for $e_i, e_j \in \mathbb{R}^d$ and u is a bijection.

The notation $\mathscr{F}(M)_x$ denotes the space of all frames at the point x. The frame bundle is then defined as

$$\mathscr{F}(M) = \bigsqcup_{x \in M} \mathscr{F}(M)_x.$$

Alongside the frame bundle, we introduce the canonical projection $\pi : \mathscr{F}(M) \to M$ by $\pi(u) = x$. The culmination of $\mathscr{F}$ and M alongside π generate a Fibre bundle with fibres $F_x = \pi^{-1}(x) = \mathscr{F}(M)_x$. Elements of $GL(n)$ act as an isomorphism on frames via composition on the right,

$$ug : \mathbb{R}^n \xrightarrow{g} \mathbb{R}^n \xrightarrow{u} T_xM.$$

The action of $GL(n)$ preserves the frames and also acts transitively on the fibres, i.e. for every $u, v \in \mathscr{F}(M)_x$, there exists a $G \in GL(n)$ s.t. $ug = v$. By using charts, can also give $\mathscr{F}(M)$ the structure of a manifold such that item iii) of Definition C.0.3 is satisfied. The result allows us to treat $(\mathscr{F}(M), \pi, M, GL(n))$ as a principle bundle. One may draw a relation between TM and $\mathscr{F}(M)$ by the action of $GL(n)$;

$$TM = \mathscr{F}(M) \times_{GL(n)} \mathbb{R}^n, \ (u, e) \mapsto ue.$$

One consequence of using elements of $GL(n)$ as a right action to elements of $\mathbb{R}^n$ is that the resulting frames are not orthonormal. To generate said orthonormal frames, we instead construct a principle bundle with the orthogonal group $O(n)$. By applying the same procedure shown above by replacing $GL(n)$ with $O(n)$, we obtain what is known as the orthonormal frame bundle. Moreover, choosing $G = O(n)$ as opposed to $GL(n)$, elements of $O(n)$ act as linear isometries from T_xM to $\mathbb{R}^n$. That is,

$$g(ua, ub) = \langle a, b \rangle_{\mathbb{R}^d}$$

for $a, b \in \mathbb{R}^n$. If for example, $\{u_i\}$ was an orthonormal basis of $\mathbb{R}^n$, generated by right action of $O(n)$, then the resulting frames $\{ue_i\}$ are also an orthonormal basis of T_xM. For the orthonormal frame bundle, we write

$$\mathcal{O}(M) = \bigsqcup_{x \in M} \mathcal{O}(M)_x$$

where $\mathcal{O}(M)_x$ is the set of orthonormal frames at x.

Again, one can show that the 4-tuple $(\mathcal{O}(M), \pi, M, O(n))$ is a principle bundle.

Example C.0.5. Recall that the canonical metric on $\mathbb{S}^2$ is $g_{\mathbb{S}^2} = d\theta^2 + \sin^2(\theta) d\phi^2$. By definition, our frame is a mapping $u : \mathbb{R}^2 \to T_{(\theta,\phi)}\mathbb{S}^2 \setminus \{S\}$ such that the resulting vector is normalised and our components are orthogonal to each other. This means that we require

$$g(ue_1, ue_1) = 1,$$

$$g(ue_1, ue_2) = 0,$$

$$g(ue_2, ue_2) = 1.$$

And so, we choose

$$u = \left\{ \partial_\theta, \frac{1}{\sin(\theta)} \partial_\phi \right\}$$

to make the mapped basis orthonormal.

With the orthonormal frame bundle constructed, we may move on to looking at its geometrical properties. With the inclusion of the Levi-Civita connection on TM, it generates an Ehresmann connection on $\mathcal{O}(M)$. This means that we can split the tangent bundle of $\mathcal{O}(M)$ into two distrinct bundles. The first bundle we look at is the vertical bundle; the set of vertical vector fields defined as the following kernel $V = \ker(d\pi : T\mathcal{O}(M) \to TM)$. Intuitively, the name vertical means that we are looking at vector fields that are orthogonal to the tangent space, and thus, in view on the tangent space, take the value of 0. The other half, the horizontal bundle, is simply the compliment of the vertical bundle. We can then write the tangent bundle of $\mathcal{O}(M)$ as a direct sum decomposition

$$T\mathcal{O}(M) = H \oplus V$$

where H is the horizontal bundle and V is the vertical bundle.

We elucidate the meaning of horizontal in the following definition:

Definition C.0.6. A curve $\{u_t\}$ on $\mathscr{F}(M)$ is called horizontal if for each $e \in \mathbb{R}^d$, the vector field $\{u_t e\}$ is parallel along $\{\pi u_t\}$; $D_{u_t e} \pi u_t = 0$. The vector field generated by $\{u_t e\}$ is called a horizontal vector field.

The curve $\{u_t\}$ is just a smooth choice of frames at each point along the curve $\{\pi u_t\}$ on M.

This new tangent space of $\mathscr{F}(M)$ can be decomposed into two distinct parts, a horizontal and a vertical part. Let $H_u\mathscr{F}(M)$ and $V_u\mathscr{F}(M)$ be the space of horizontal and vertical vectors at u; then we have the following decomposition

$$T_u\mathscr{F}(M) = V_u\mathscr{F}(M) \oplus H_u\mathscr{F}(M).$$

Since the canonical projection $\pi : \mathscr{F}(M) \to M$ induces an isomorphism on the horizontal tangent space $\pi_* : H_u\mathscr{F}(M) \to T_{\pi u}M$ there is always a unique horizontal vector field for every vector field on TM. For any vector $v \in T_{\pi(x)}M$, there exists a unique horizontal vector $v_x^* \in H_x$ such that $d\pi(v_x^*) = v$

For each $e \in \mathbb{R}^n$, the associated horizontal vector field H_e on $T\mathcal{O}(M)$ defined at $u \in \mathcal{O}(M)$ is defined by

$$H_e(u) = (ue)^*.$$

This is known as the horizontal lift of the vector ue. Using the relation above, $d\pi(H_e) = ue$ with $ue \in T_{\pi u}M$. When $\{e_i\}$ are an orthonormal basis of $\mathbb{R}^n$, the resulting horizontal vector fields are called the fundamental horizontal vector fields; $H_{e_i} = (ue_i)^*$. Furthermore, these vector fields span the horizontal tangent space H_u at u.

Appendix D

Brownian Motion on Lie Groups

In this section, let G be a Lie group with Lie algebra $\mathfrak{g}$ of dimension d.

Definition D.0.1. A Process $\{X\}_{t\in\mathbb{R}_+}$ with values in G is called a left-invariant Brownian motion on G if:

i) $\{X\}_{t\in\mathbb{R}+}$ is continuous,

ii) for each $s \geq 0$, the process $\{X_s^{-1}X_{s+t} : t \geq 0\}$ is independent of the process $\{X_r : r \leq s\}$,

iii) for each $s \geq 0$ the processes $\{X_s^{-1}X_{s+t} : t \geq 0\}$ and $\{X_t : t \geq 0\}$ are identical in law.

Each left-invariant Brownian motion on G is a Feller-Dynkin diffusion. Its transition function is specified by the action of its generator $\mathcal{L}$ on $C_c^{\infty}(G)$,

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{d} U_i^2 + V$$

where U_i, $(i = 1, ...d)$ and V are both LIVF's.

Given $A \in \mathfrak{g}$ is an orthonormal basis element of $\mathfrak{g}$, the associated vector field U is a first order operator on smooth functions f on G defined by

$$U(f(X)) = \frac{d}{dt} f(Ge^{tA})\bigg|_{t=0}$$

Suppose we have the generator $\mathcal{L}$ of a diffusion on G, $\{X\}_{t\in\mathbb{R}_+}$. Let A_i and C be the LIVF's in $\mathfrak{g}$ associated with the vector fields U_i and V in the generator $\mathcal{L}$ respectively. Let $\{B_t\}_{t\in\mathbb{R}^+}$ be a standard Brownian motion on $\mathbb{R}^d$ and let

$$A_t = B_t^q A_q + tC.$$

Then $\{A_t\}_{t\in\mathbb{R}_+}$ is a (left) Brownian motion on $\mathfrak{g}$. Using the Brownian motion on $\mathfrak{g}$, we may construct a Brownian motion on the Lie group G. Let $X \in G$ and define $\{X_t\}_{t\in\mathbb{R}_+}$ be the solution to the SDE

$$dX_t = X_t \circ dA_t, \quad X_0 = X.$$

Then $X_t \in G$ and is also a (left) Brownian motion on G with generator $\mathcal{L}$.

Example D.0.2. Let $G = \mathrm{SO}(3)$ and so $\mathfrak{g} = \mathfrak{so}(3) = \mathrm{skew}(3)$. Define

$$A = A(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{pmatrix}$$

for $x = (x_1, x_2, x_3) \in \mathbb{R}^3$. Then if $\{B_t\}_{t\in\mathbb{R}_+}$ is a Brownian motion in $\mathbb{R}^3$ with independent components B^1, B^2, B^3, $\{A(B)t\}_{t\in\mathbb{R}_+}$ is a canonical Brownian motion on $\mathfrak{so}(3)$. We can verify this by checking that the quadratic variation process is simply Idt; since we have that

$$dA_t = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & -dB_t^3 & dB_t^2 \\ dB_t^3 & 0 & -dB_t^1 \\ -dB_t^2 & dB_t^1 & 0 \end{pmatrix}$$

an exercise in matrix algebra yields $d[A]_t = dA_t dA_t^\intercal = Idt$. Now, let $S \in \mathrm{SO}(3)$. Then the process $\{S_t\}_{t\in\mathbb{R}_+}$ is a Brownian motion on SO(3) which solves the

Stratonovich SDE

$$dS_t = S_t \circ dA_t.$$

Via the Itô-Stratonovich conversion formula, we can transform to Itô form, resulting in

$$\begin{aligned} dS_t &= S_t dA_t + \frac{1}{2} dS_t dA_t \\ &= (S_t + \frac{1}{2} dS_t) dA_t. \end{aligned}$$

In order to isolate dS_t on the left hand side, we require some further manipulation;

$$dS_t\left(I - \frac{1}{2}dA_t\right) = S_t dA_t,$$

$$dS_t = S_t dA_t \left(I - \frac{1}{2}dA_t\right)^{-1}.$$

Note that $I - \frac{1}{2}dA_t$ is non-singular and so the matrix inverse does exist. The next step we take is to explicitly calculate this inverse term;

$$\left(I - \frac{1}{2}dA_t\right)^{-1} = \frac{1}{8+3dt}\begin{pmatrix} 8+dt & -2\sqrt{2}dB_t^3 & 2\sqrt{2}dB_t^2 \\ 2\sqrt{2}dB_t^3 & 8+dt & -2\sqrt{2}dB_t^1 \\ -2\sqrt{2}dB_t^2 & 2\sqrt{2}dB_t^1 & 8+dt \end{pmatrix}.$$

Now since $|\frac{3}{8}dt| \ll 1$ we may apply the geometric series formula whilst ignoring all 2nd order and higher terms so that

$$\frac{1}{8+3dt} = \frac{1}{8} - \frac{3}{64}dt.$$

Therefore we reduce the inverse down to a much simpler form,

$$\left(I - \frac{1}{2}dA_t\right)^{-1} = \frac{1}{8}\begin{pmatrix} 8+dt & -2\sqrt{2}dB_t^3 & 2\sqrt{2}dB_t^2 \\ 2\sqrt{2}dB_t^3 & 8+dt & -2\sqrt{2}dB_t^1 \\ -2\sqrt{2}dB_t^2 & 2\sqrt{2}dB_t^1 & 8+dt \end{pmatrix} - \frac{3}{8}Idt.$$

And finally, some further matrix algebra allows us to show that

$$dA_t\left(I - \frac{1}{2}dA_t\right)^{-1} = \begin{pmatrix} -\frac{1}{2}dt & -\frac{1}{\sqrt{2}}dB_t^3 & \frac{1}{\sqrt{2}}dB_t^2 \\ \frac{1}{\sqrt{2}}dB_t^3 & -\frac{1}{2}dt & -\frac{1}{\sqrt{2}}dB_t^1 \\ -\frac{1}{\sqrt{2}}dB_t^2 & \frac{1}{\sqrt{2}}dB_t^1 & \frac{1}{2}dt \end{pmatrix}$$

which we can express more neatly as

$$dA_t\left(I - \frac{1}{2}dA_t\right)^{-1} = dA_t - \frac{1}{2}Idt.$$

Whence, our Brownian motion on SO(3) obeys the following SDE:

$$dS_t = S_t dA_t - \frac{1}{2}S_t dt.$$

It is now possible to verify two things: First, that S_t is indeed in SO(3), and secondly, the quadratic variation is Idt. To check the first, we examine the process $d(S_tS_t^\intercal)$, where $dS_t^\intercal = dA_t^\intercal S_t^\intercal - \frac{1}{2}S_t^\intercal dt$. First, we note $dS_tdS_t^\intercal = S_tdA_tdA_t^\intercal S_t^\intercal = S_tS_t^\intercal dt$, then

$$\begin{aligned} d(S_tS_t^\intercal) &= dS_tS_t^\intercal + S_tdS_t^\intercal + dS_tdS_t^\intercal \\ &= S_tdA_tS_t^\intercal - \frac{1}{2}S_tS_t^\intercal dt + S_tdA_t^\intercal S_t^\intercal - \frac{1}{2}S_tS_t^\intercal dt + dS_tdS_t^\intercal \\ &= -S_tS_t^\intercal dt + S_tS_t^\intercal dt \\ &= 0. \end{aligned}$$

Therefore, since $S_0 = S \in \mathrm{SO}(3)$, integration yields $S_tS_t^\intercal = SS^\intercal = I$. Moreover, we can go back to the variation formula and conclude that $dS_tdS_t^\intercal = Idt$.