Stockholms universitet

Geometrically Aware Markov Chain Monte Carlo

Hedwig Nora Nordlinder

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2025:18
Matematisk statistik
Juni 2025

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2025:18**
http://www.math.su.se

Geometrically Aware markov Chain Monte Carlo

Hedwig Nora Nordlinder*

June 2025

Abstract

In this thesis we study the generalisation of the Metropolis Adjusted Langevin Algorithm to the Riemannian manifold of symmetric positive definite matrices P(n). Specifically, an application to hierarchical models that involve the Wishart distribution are considered. A concrete example is given for modelling the rates of synonymous and non-synonymous substitution in a phylogeny. It is proven that a large class of uniformly log-concave posterior densities attain -bounded Wasserstein distance from their invariant measures in O(2) iterations of the Riemannian Metropolis-adjusted Langevin Algorithm. It is also shown that common generalisations of the LKJ-distribution never satisfy a set of sufficient conditions for this bound. Lastly, it is conjectured that certain conditions for attaining the iteration complexity bound may be weakened to hold probabilistically only.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: hedwignordlinder@gmail.com. Supervisor: Ola Hössjer, Johannes Heiny, Benjamin Murrell.

Acknowledgements

First and foremost, I am very grateful to my supervisors Prof. Ola Hössjer, Dr. Johannes Heiny, Dr. Benjamin Murell. It probably goes without saying that this thesis would not have been possible without their help in fields from phylogenetics to matrix inequalities. I also owe a great deal of gratitude to the large language models Claude and Grok, which have helped me extensively when writing code, searching for references and generating figures. After these, the most important contribution to the completion of this thesis comes from the provision of both practical and emotional support from my parents. Lastly, I extend a special thank you to amanuensis Leo G. Levenius for encouraging (and persuading) me to pursue a BSc in mathematical statistics, preventing me from going down the stray path of an engineering degree, as well as proof reading the thesis.

Contents

Chapter 1

Motivating Problem in Phylogenetics

> Natural selection is a mechanism for generating an exceedingly high degree of improbability.
>
> — Sir R. A. Fisher

The problem of studying the evolutionary history of a set of organisms that may or may not have evolved from a common ancestor is known as *phylogenetic inference*. The motivating practical problem for the results presented in this thesis specifically relates to performing phylogenetic inference at the molecular (more specifically codon) level. Before introducing the problem at hand, a brief introduction to (codon) molecular evolution will follow.

1.1 Codon Molecular Evolution Models

The human genome is encoded as DNA, which is a macromolecule consisting of two strands coiled together. These strands in turn consist of *nucleotides*. The exact chemical structure of nucleotides is largely immaterial to phylogenetic statistical models, so for simplicities sake we (erroneously) identify nucleotides only with their corresponding nitrogenous base. There are four nitrogenous bases present in DNA: Adenine [A], Cytosine [C], Guanine [G] and Thymine [T]. Triplets of these nitrogenous base pairs are called *codons*, which encode information about what amino acid should be produced by ribosomes during translation. Ultimately, these amino acids bind together to form proteins. There are $4^3 = 64$ possible triplets of nitrogeneous bases (codons) but only 20 proteinogenic amino acids. Some codons are what are known as *stop codons*, which tell the ribosome to stop transcription, but there are also duplicate codons, that is several different codons that code for the same amino acid (see Figure 1.1)

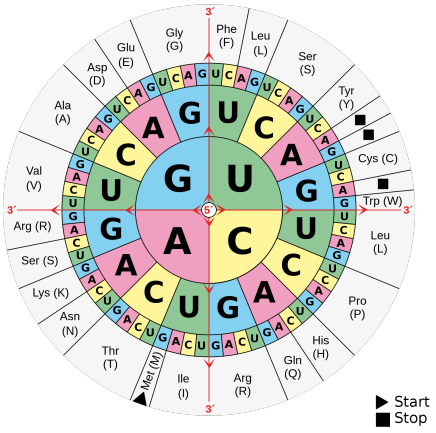

Figure 1.1: Transcription wheel mRNA → amino acids [10]

.

One can study the evolutionary history of a set of organisms by studying how their DNA differs. More specifically one can consider what codons are present at specific *sites* in their genetic code. By studying the (dis)similarity of organisms at various sites in their DNA one can build a *phylogenetic tree*. An example of a phylogenetic tree organising Darwin's finches is given in Figure 1.2. This tree is what is known as a *ethological* phylogenetic tree, where species are grouped by their behavioural characteristics (ethological traits). In the example provided, these ethological traits are the hunting patterns exhibited by the different subspecies.

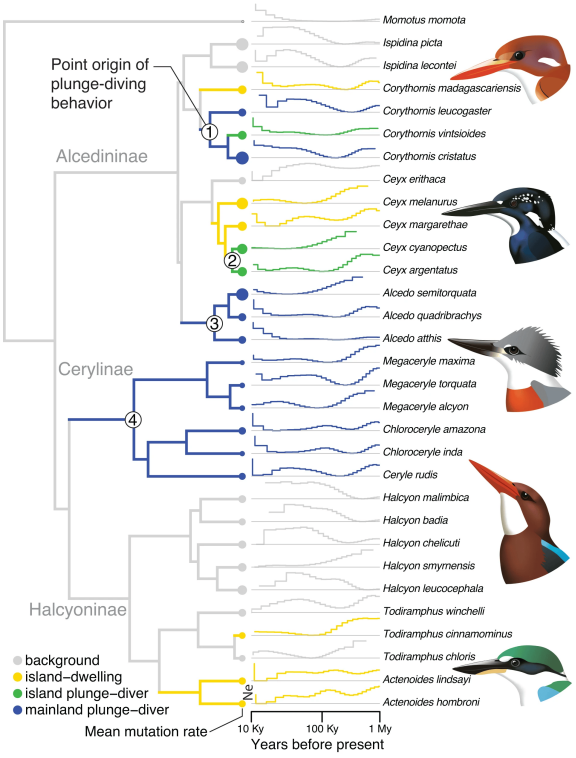

Figure 1.2: A phylogenetic tree of Darwin's finches, [4] [1]

Trees such as the one shown in Figure 1.2 provide a pedagogic example, but molecular evolution

[1] Reproduced under the Creative Commons Attribution 4.0 International License

based methods for inferring phylogenetic trees instead construct the tree topology by maximum likelihood[2] under some model of molecular evolution. More formally let τ denote the *topology* of a phylogenetic tree, let θ be parameters of some model of molecular evolution and let $\mathcal{D}$ be the DNA sequences of the organisms we wish to construct a phylogenetic tree over. Then the tree topology $\hat{\tau}$ that satisfies

$$(\hat{\tau}, \hat{\theta}) := \arg\max_{\tau,\theta} \mathbb{P}(\mathcal{D}|\tau, \theta)$$

is called the maximum likelihood estimated tree for the data $\mathcal{D}$.

To understand how such likelihood calculations are performed we need to understand *codon substitution models*. A codon substitution model is a model of molecular evolution that describes changes in nitrogenous base pairs at a site as the realisation of a continuos time Markov chain (CTMC). One such model is the generalised time reversible (GTR) model. If we let $(\pi_A, \pi_C, \pi_G, \pi_T)$ denote the equilibrium frequencies of the nitrogenous bases A, C, G, T. Under the GTR model we estimate exchangeability parameters $r_{AC}, r_{AG}, r_{AT}, r_{CG}, r_{CT}, r_{GT}$. From this the transition rate matrix of the continuos time Markov chain can be constructed as

$$Q := \begin{pmatrix} * & \pi_C r_{AC} & \pi_G r_{AG} & \pi_T r_{AT} \\ \pi_A r_{AC} & * & \pi_G r_{CG} & \pi_T r_{CT} \\ \pi_A r_{AG} & \pi_C r_{CG} & * & \pi_T r_{GT} \\ \pi_A r_{AT} & \pi_C r_{CT} & \pi_G r_{GT} & * \end{pmatrix}$$

where the diagonal elements are chosen so that the row-sums are 0, as required for a transition rate matrix.
To calculate the likelihood of a tree we also need some specification of the transition rates between codons, not just nucleotide bases. While several such models exist we will choose to describe the MG94 model [12] to provide conceptual understanding. The MG94 model defines a 61×61 matrix for transitions between the 61 non-stopping codons as

$$M_{ij,i\neq j} = \begin{cases} 0, \text{ if codons } i \text{ and } j \text{ differ by more than one nitrogeneous base} \\ \alpha \times q_{\text{origin codon, target codon}} \text{ if the substitution is synonymous} \\ \beta \times q_{\text{origin codon, target codon}} \text{ if the substitution is non-synonymous} \end{cases}$$

where q_{ij} are entries of the GTR rate matrix. Once again the diagonal entries are chosen so that the rows sum to zero. In this model α and β denote the rates of synonymous and non-synonymous substitution, which are parameters of our codon model. By standard theory for stochastic processes this means that we can compute a transition probability of a transition happening over some evolutionary distance t as

$$P_{ij}(t) = e^{tM}{}_{ij}.$$

Here the evolutionary distance t is given by the *branch length* of the tree, which is an inferred parameter of our model.

Under the models described the likelihood of a combination of tree parameters and codon model parameters can be computed as

$$p(\mathcal{D}|\mathcal{T}, Q, \theta) = \prod_k \int_{\mathbb{R}^2} p(\mathcal{D}_k|\alpha, \beta, \mathcal{T}) p(\alpha, \beta|\theta) d\alpha d\beta$$

by simple marginalisation argument. Here $\mathcal{D}_k$ denotes the sequence data for the genetic site k (assumed to be conditionally independent) and $p(\alpha, \beta|\theta)$ denotes a parametrised bivariate distribution over synonymous and non-synonymous substitution rates. It should not come as a surprise that the integral is intractable. We will therefore proceed by discretising the underlying continuos bivariate distribution. One such method of discretisation is provided by the FUBAR method, which is described in the next section

[2]There are other methods, such as maximum a posteriori estimation under a Bayesian model or maximum parsimony methods, but our focus will be on maximum likelihood inferred trees

1.1.1 FUBAR

The FUBAR method [11] discretises the α, β-distribution by choosing $p(\alpha, \beta|\theta)$ to be the general bivariate distribution over a finite set of synonymous and non-synonymous rates. θ is therefore simply a probability vector, yielding a model with minimal parametric assumptions. The likelihood function therefore simplifies significantly to

$$p(\mathcal{D}|\mathcal{T}, Q, \theta) = \prod_k \sum_{\alpha,\beta} p(\mathcal{D}_k|\alpha, \beta, \mathcal{T}) p(\alpha, \beta|\theta)$$

(refer to equation (2) in [11]). In the original FUBAR paper θ is inferred under the Bayesian framework where the prior distribution over θ is chosen to be the Dirichlet distribution, which is supported on the set of probability vectors, that is the set defined as $\{x \in \mathbb{R}^n : x_i \geq 0, \sum_i x_i = 1\}$. One can conceptually think of the Dirichlet distribution as a multivariate generalisation of the Beta distribution. Just as the Beta distribution is the conjugate prior for a binomial likelihood, the Dirichlet distribution is the conjugate prior for a multinomial likelihood. When data is sparse, this prior distribution can yield high posterior probability estimates to nonsensical discretisations of an underlying continuos distribution. A method currently under development by the author is SKBDI (Smooth Kernel Based Density Inference). The problems associated with FUBAR and their resolution by SKBDI will be described below

1.1.2 SKBDI

The core problem with the Dirichlet prior used in FUBAR is that it does not assign a low prior probability to "spiky" distributions, that is distributions where neighbouring points on a grid of synonymous and non-synonymous substitution have vastly different probabilities. Intuitively since θ represents a discretisation of an underlying distribution that is assumed to be continous it should a priori be very unlikely to see such sharp differences in probability assigned to neighbouring points. It is, of course, theoretically possible that the posterior probability of such a discretisation is high, but this should require strong evidence from the likelihood function. When the data is weakly informative, we wish to avoid such spikiness. To better illustrate how this problem arises, consider the following coin tossing example:

We have a sample of n coins, and denote the probability of tails for coin i by p_i. We assume these probabilities were drawn from a common continuos distribution supported on $[0, 1]$, which we wish to infer from our data $\mathcal{D}$. Here our data is a record of the amount of times each coin yielded tails, after being tossed 100 times. Applying a FUBAR-like method would mean that we discretise our distribution to be some amount of categories of probabilities, perhaps $[0, 0.01), [0.01, 0.02) \cdots [0.99, 1]$ and use a Multinomial-Dirichlet model to find a posterior distribution over these categories. Figure 1.3 shows the aforementioned spikiness artifact when this discretisation is applied to approximate a "true" Beta$(5, 5)$ distribution. As one can see the problem is negligible when data (amount of tosses) is abundant, but can cause significant problems when data is sparse.

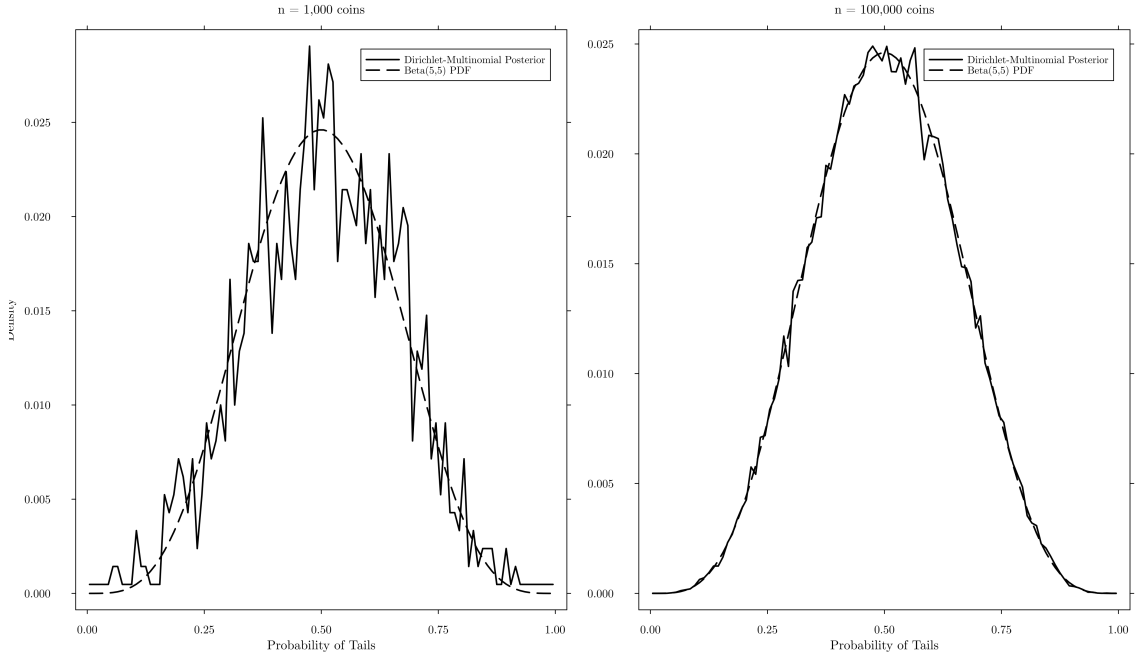

Figure 1.3: Discretisation error caused by Dirichlet-Multinomial approximation

This problem motivates us to somehow encode in our prior an enforcement of smoothness (which can be overruled by the likelihood function the data supports a discontinuity). SKBDI does this by defining a hierarchical model in the following way[3]

$$\log c \sim \mathcal{N}(0, \sigma_0),$$
$$\log \theta \sim \mathcal{N}(\mathbf{0}, \Sigma(c)).$$

where we define

$$\Sigma(c)_{ij} := ke^{\frac{-d(i,j)^2}{e^c}}$$

that is, we generate $\Sigma(c)$ using a Gaussian kernel function. Furthermore, θ is sampled in ambient Euclidean space and transformed by the softmax operation[4] to be a probability vector. We choose to parametrise the covariance matrix with c since the problem of inferring the full covariance matrix would be severely underdetermined for only one alignment. To allow us to perform inference on the full covariance matrix without this strong parametric assumption we need *several different alignments*. Intuitively, this is similar to what one learns in an introductory statistics course: It is not possible to estimate the variance from a single sample. We therefore introduce the following model:

$$\Sigma \sim \mathcal{D},$$
$$\log \theta_i \sim \mathcal{N}(\mathbf{0}, \Sigma).$$

where $\theta_1, \theta_2 \cdots \theta_n$ are grid estimates for different alignments and $\mathcal{D}$ is some arbitrary distribution supported on $\mathcal{P}(n)$. In Chapter 4 it is shown that the *Riemannian Gaussian* distribution is the most theoretically sound prior, but practical results show that the Wishart distribution also works well.

[3]The current implementation uses parametric kernel matrices and does not sample over the full space of symmetric positive definite matrices.

[4]An interesting reflection is that we are using a non translation invariant prior for a translation invariant problem, which is suboptimal.

Chapter 2

Background in Differential Geometry

> Geometry is the art of correct reasoning from incorrectly drawn figures.
>
> — Henri Poincaré

Here we will give the reader a gentle repetition of differential geometry required to understand a theoretical result given in chapter 4. Readers wishing to fill in gaps left by the repetition given here are encouraged to refer to [9].

2.1 Basic definitions

Definition 2.1.1. (Chart)
A *chart* (U, φ) on a topological space M is an open subset $U \subset M$ together with a *homeomorphism* $\varphi : U \to S \subset \mathbb{R}^n$ such that S is open.

Definition 2.1.2. (Atlas)
An *atlas* for a topological space M is an indexed family $\{(U_\alpha, \phi_\alpha) : \alpha \in I\}$ of charts on M such that the charts *cover* M ($\bigcup_{\alpha \in I} U_\alpha = M$). An atlas $\mathcal{A}$ is said to be *maximal* if there does not exist any atlas $\mathcal{B}$ such that $\mathcal{A} \subset \mathcal{B}$

Definition 2.1.3. (Transition map)
Let $\mathcal{A}$ be an atlas and let $(U_\alpha, \varphi_\alpha), (U_\beta, \varphi_\beta) \in \mathcal{A}$ be charts such that $U_\alpha \cap U_\beta \neq \emptyset$. The *transition map* $\tau_{\alpha,\beta} : \varphi_\alpha(U_\alpha \cap U_\beta) \to \varphi_\beta(U_\alpha \cap U_\beta)$ is the map defined by $\tau_{\alpha,\beta} = \varphi_\beta \circ \varphi_\alpha^{-1}$

Definition 2.1.4. (Differentiable atlas)
A *differentiable atlas* is an atlas where all transition maps are differentiable functions.

With this background we are now ready to give a central definition for this paper:

Definition 2.1.5. (Differentiable manifold)
A *differentiable manifold* is a topological space[1] M together with a maximal differentiable atlas on M

The notion of a differentiable manifold allows us to essentially "perform calculus" in non-traditional settings where the set of interest can be endowed with the structure of a differentiable manifold. We say that a function $f : M \to \mathbb{R}$ is differentiable at $p \in M$ if and only if for some differentiable chart (U, φ) with $p \in U$ $f \circ \varphi^{-1} : \varphi(U) \subset \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\varphi(p)$ in the traditional sense

[1] Formally, we require the topological space to be second countable and Hausdorff. We avoid mentioning this in the main text since it adds verbosity but not intuition, nor is it used in later sections

of differentiability in $\mathbb{R}^n$. Differentiable manifolds are not necessarily vector spaces so the notion of the directional derivative of a function needs to be defined along a differentiable curve γ on M with $\gamma(0) = p$. The directional derivative of f at p along γ is then defined as $\frac{d}{dt} f(\gamma(t))\Big|_{t=0}$. We are now ready to give another central definition:

Definition 2.1.6. (Tangent Vector Space)
We can define an equivalence relation on the set of curves over M as follows: If γ_1 and γ_2 are two curves with $\gamma_1(0) = \gamma_2(0) = p$ satisfying that $\frac{d}{dt}\varphi \circ \gamma_1(t)\Big|_{t=0} = \frac{d}{dt}\varphi \circ \gamma_2(t)\Big|_{t=0}$ for every chart φ then $\gamma_1 R \gamma_2$. The set of all curves passing through p at 0 modulo this equivalence relation defines a vector space known as the tangent vector space at p, denoted by T_pM.

The operations on this vector space are defined by the mapping $d\varphi_p : T_pM \to \mathbb{R}^n$ by $d\varphi_p(\gamma'(0)) := \frac{d}{dt}\varphi \circ \gamma(t)\Big|_{t=0}$ where $\gamma'(0) \in T_pM$, γ is some member of the equivalence class represented by $\gamma'(0)$ and φ is some chart. The resulting vector space operations are independent of the choice of chart and the mapping $d\varphi_p$ turns out to be a bijection. Figure 2.1 shows the tangent space of the sphere at a point p, "sticking out" into the ambient Euclidean space $\mathbb{R}^3$ If we endow the tangent space

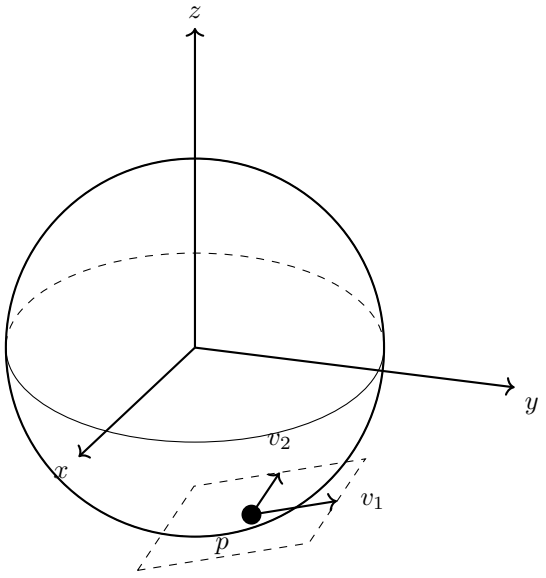

Figure 2.1: The 2-sphere S^2 in $\mathbb{R}^3$ with the tangent space T_pS^2 at a point p.

$T_p\mathcal{M}$ with an inner product our differentiable manifold becomes a *Riemannian manifold*, which will be the central object of study going forward. As we recall from undergraduate linear algebra inner products induce norms.

2.2 Riemannian Manifolds

If we have an inner product $\langle \cdot, \cdot \rangle : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$ we can define the norm of a vector v by $||v||^2 := \langle v, v \rangle$. This also gives rise to a notion of distance. But this holds on the tangent space only, but one can use this definition to define the distance between two points p, q on a manifold in the following way

Definition 2.2.1. (Distance between two points on a Riemannian manifold) Let

$$d(p, q) := \inf_{\gamma} \int_0^1 ||\gamma'(t)|| dt,$$

where γ is a differentiable curve $\gamma : [0, 1] \to \mathcal{M}$ satisfying $\gamma(0) = p$ and $\gamma(1) = q$. Therefore we define the distance between points as the minimum curve length of a differentiable curve connecting them.

If we take these differentiable curves to have constant speed, that is $||\gamma'(t)|| = c$ we obtain what are called *geodesics*. Geodesics are the Riemannian generalisation of straight lines in Euclidean space. The core intuition used to describe differentiable manifolds is to imagine Euclidean space, but replacing the familiar straight lines geodesics. This is what is meant by Riemannian geometry being curved. Before discussing curvature in more detail we will define the *exponential* and *logarithmic* maps.

Definition 2.2.2. (The exponential map) Let $v \in T_p\mathcal{M}$ be a tangent vector of the point p and consider the geodesic $\gamma_v : [0,1] \to \mathcal{M}$ satisfying $\gamma_v(0) = p$ and $\gamma_v'(0) = v$[2]. We define the exponential map to be

$$\text{Exp}_p(v) := \gamma_v(1).$$

and call its inverse the logarithmic map.

The name "exponential map" comes from the fact that the mapping with this property on the Riemannian manifold of positive real numbers is defined as $\exp_p(v) = pe^v$. The tangent vector space of this manifold is isomorphic to $\mathbb{R}$. A very useful informal conceptualisation is to think of the exponential map as saying "What if we want to perturb the point p by a small amount v". This intuition is especially useful when defining stochastic processes on differentiable manifolds, since we do not normally have a well-defined notion of addition of distinct points. To further understand the theory of stochastic processes on differentiable manifolds we will dedicate the next section to discussing the *Riemannian curvature tensor*, which allows measurement of the curvature of a manifold.

2.2.1 Curvature

We will not give a formal definition of the Riemannian curvature tensor, as it would require too much background in subjects less relevant to the core goal of this thesis. Instead, we choose an intuitive description which will first require us to recall the definition of a *vector field*.

Definition 2.2.3. (Vector field on a Riemannian manifold). Assign to each point p on the manifold $\mathcal{M}$ an element of the tangent vector space $T_p\mathcal{M}$. This is a vector field on $\mathcal{M}$.

Most readers will be familiar with one type of vector field: Consider the Riemannian manifold $\mathbb{R}^d$ (that is, regular Euclidean space) and let $f : \mathbb{R}^d \to \mathbb{R}$ be some differentiable function. Associate with each point $p \in \mathbb{R}^d$ the gradient of the function f at the point p, that is $\nabla f(p)$. Since the gradient of f evaluated at a point is a d-dimensional vector this means we have associated with each point on $\mathbb{R}^d$ an element of the tangent space of $\mathbb{R}^d$ (which is just $\mathbb{R}^d$ again). Thus we have defined a vector field on $\mathbb{R}^d$ known as the *gradient vector field*.

Now let us consider three vector fields X Y and Z defined on the Euclidean space. At the point $p \in \mathbb{R}^d$ we obtain two vectors $X(p)$ and $Y(p)$. These two vectors span a parallelogram. If we first transport the vector $Z(p)$ a small distance in the direction of $X(p)$ and then transport it a small distance in the direction $Y(p)$ the result is the same as if we had done this operation in the reverse order, as shown in Figure 2.2. Furthermore, if transport $Z(p)$ go "all the way around" the parallelogram we get back to the same place we started

[2]Since the geodesics can be defined as solutions to differential equations we can use the existence and uniquness properties of these to claim that γ_v is uniquely determined by these two conditions

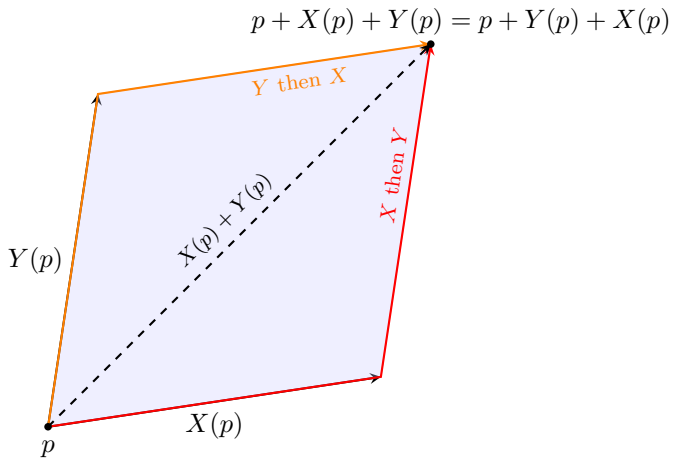

Figure 2.2: Parallel transport of $Z(p)$ along $X(p)$ and $Y(p)$ in $\mathbb{R}^d$.

We are now ready to give an intuition for what the Riemannian curvature tensor is and what it measures. Let $\mathfrak{X}(\mathcal{M})$ denote the set of all possible vector fields on $\mathcal{M}$. The Riemannian curvature tensor is a mapping $R : \mathfrak{X}(\mathcal{M}) \times \mathfrak{X}(\mathcal{M}) \times \mathfrak{X}(\mathcal{M}) \to \mathfrak{X}(\mathcal{M})$ that measures the failure of this operation to behave as it does in Euclidean space (that is, transportation along an infinitesimal parallelogram taking us back to the point we started at). Precisely, the Riemannian curvature tensor gives us a vector field of vectors that measure the difference between the starting point and the result of transportation around these infinitesimal parallelograms. This failure to return to the starting point is exactly what is meant by non-Euclidean geometries having curvature, and the Riemannian curvature tensor measures exactly this curvature. Using it, we can define two other notions of curvature: *Sectional curvature* and the *Ricci curvature*

Definition 2.2.4. (Sectional curvature of a Riemannian manifold) Let $u, v \in T_p\mathcal{M}$ be two non-parallel tangent vectors at a point $p \in \mathcal{M}$ and let

$$K(u, v) := \frac{\langle R(u, v)v, u\rangle}{\langle u, u\rangle\langle v, v\rangle - \langle u, v\rangle^2}.$$

K is called the *sectional curvature* of $\mathcal{M}$ at the point p.

Since we require u, v to be non-parallel the expression in the denominator will be strictly positive by the Cauchy-Schwartz inequality, meaning the expression is well-defined. Furthermore u, v will span a two-dimensional subspace of the tangent space $T_p\mathcal{M}$. The sectional curvature can therefore be thought as defining how curved space is along this two-dimensional plane. Lastly, we will define the Ricci curvature

Definition 2.2.5. (Ricci curvature) The mapping $\mathrm{Ric}_p : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$ defined as

$$\mathrm{Ric}_p(Y, Z) := \mathrm{Tr}(X \mapsto R_p(X, Y)Z)$$

is called the Ricci curvature tensor

To clarify, the mapping $X \mapsto R_p(X, Y)Z$ is a linear mapping of the tangent space $T_p\mathcal{M}$ to itself, since the Riemannian curvature tensor at p has the set of vector fields on $\mathcal{M}$ as its image, and here we are considering $R_p(X, Y)Z$ to be the vector field obtained from the Riemannian curvature evaluated at p, hence it is a tangent vector in the tangent space of the point p. Since this is a linear mapping of finite-dimensional vector spaces[3] it admits a representation as a matrix, and therefore the trace of it is well-defined. For some intuition, one can consider a collection of geodesics starting close to the point p, initially pointing in directions close to v. Then the sign of $\mathrm{Ric}_p(v, v)$ will determine if these geodesics will converge, remain parallel or diverge (see Figure 2.3)

[3] A linear endomorphism on $T_p\mathcal{M}$, if one wishes

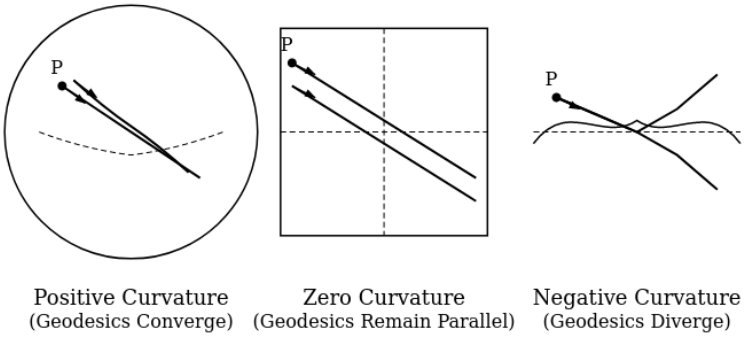

Figure 2.3: Behaviour of geodesics under various curvatures

Relevance to stochastic processes

So far, a lot of mathematical objects relating to the curvature of Riemannian manifolds have been defined, but it is not immediately clear why they are relevant to understanding the behaviour of stochastic processes on manifolds. Recall that the Metropolis-adjusted Langevin Algorithm is based on a Wiener process, which is a type of *diffusion process*. One intuition for stochastic processes on manifolds is to think of a probability distribution as a source of heat, and to think of stochastic processes governed by this probability distribution as analogous to heat diffusion along the manifold. If geodesics tend to naturally converge (as happens when the Ricci curvature is positive) then the distribution of heat on the manifold will be more "peaked" and stochastic processes governed by this "heat source" will have a naturally contractive behaviour, staying in regions of high probability. In the case of negative curvature however this natural contraction property is not present, and the heat distribution will be more diffuse.

2.2.2 The Manifold $\mathcal{P}(n)$ with Affine-Invariant Metric

We will now proceed to give some definitions specifically relating to the manifold $\mathcal{P}(n)$ when its tangent space is endowed with the Affine Invariant Riemannian Metric. Recall that $\mathcal{P}(n)$ is the set of all symmetric positive definite matrices, that is $\mathcal{P}(n) := \{X \in M_n(\mathbb{R}) : u^t X u \geq 0 \ \forall u \in \mathbb{R}^n \setminus \{0\}\}$

Definition 2.2.6. (Tangent Space of $\mathcal{P}(n)$). The tangent space of $\mathcal{P}(n)$ is simply the set of symmetric matrices, that is all matricies X with the property $X = X^T$

Definition 2.2.7. (The Affine Invariant Riemannian Metric on $\mathcal{P}(n)$) Let $\langle \cdot, \cdot \rangle_p : T_p\mathcal{P}(n) \times T_p\mathcal{P}(n) \to \mathbb{R}$ be defined as

$$\langle A, B \rangle_p := \mathrm{Tr}(p^{-1}Ap^{-1}B).$$

This defines the Affine Invariant Riemannian Metric on the tangent space of $\mathcal{P}(n)$ [14]. Affine invariance in this context means that the metric is invariant under the action of the orthogonal group on $\mathcal{P}(n)$ by conjugation.

We will now recall the geodesics, exponential map and logarithmic map as well as the parallel transport maps, all being central to understanding the material that will follow

Definition 2.2.8. (The Exponential map on $\mathcal{P}(n)$). The mapping $\mathrm{Exp}_X(Y) : T_X\mathcal{P}(n) \to \mathcal{P}(n)$ given by

$$\mathrm{Exp}_X(Y) = X^{\frac{1}{2}} \exp\left(X^{-\frac{1}{2}} Y X^{-\frac{1}{2}}\right) X^{\frac{1}{2}}$$

is the exponential map on $\mathcal{P}(n)$

Definition 2.2.9. (The geodesics on $\mathcal{P}(n)$). The geodesics on $\mathcal{P}(n)$ are given by

$$\gamma(t) = X^{\frac{1}{2}} \exp\left(t \log\left(X^{-\frac{1}{2}} Y X^{-\frac{1}{2}}\right)\right) X^{\frac{1}{2}}$$

Curvature on $\mathcal{P}(n)$

The Riemannian manifold $\mathcal{P}(n)$ is what is known as a *Cartan-Hadamard* manifold. The exact definition is somewhat involved and not relevant, but a very central property is that such a manifold has everywhere negative sectional curvature. This gives rise to complications for diffusion processes defined on $\mathcal{P}(n)$ since we do not get contractivity "for free" from the geometry. This gives rise to what in [3] is called a "harder but more general setting". It is more involved to prove useful properties of certain diffusion properties when we have negative curvature, which aligns with our earlier intuition that the divergence of almost parallel geodesics starting close to each other is an undesirable property for stochastic processes.

Chapter 3

Markov Chain Monte Carlo

> The past is but the beginning of a beginning, and all that is or has been is but the twilight of the dawn.
>
> —H.G. Wells

3.1 Statistical problem

We will focus on the following statistical problem: We wish to know the probability distribution of a set of parameters θ, Σ where we believe a-priori that $\theta \sim N(\mu_0, \Sigma),\ \Sigma \sim \mathrm{W_k}(\Sigma_0, n_0)$ with $W_k(\Sigma_0, n_0)$ denoting the matrix-variate *Wishart distribution*, which is supported on $\mathcal{P}(k)$. We will denote our data by $\mathbf{x}$, our likelihood by $L(\mathbf{x}|\theta, \Sigma)$ and our prior density by $\pi(\theta, \Sigma)$. By Bayes theorem this gives us the posterior distribution for our parameters

$$\pi(\theta, \Sigma|\mathbf{x}) = \frac{L(\mathbf{x}|\theta, \Sigma)\pi(\theta, \Sigma)}{\int_\Theta L(\mathbf{x}|\theta, \Sigma)\pi(\theta, \Sigma)d\Theta} = \frac{L(\mathbf{x}|\theta, \Sigma)\pi(\theta, \Sigma)}{\pi(\mathbf{x})} \propto \pi(\mathbf{x}|\theta, \Sigma)\pi(\theta, \Sigma).$$

where Θ denotes our parameter space. For some problems, the integral in the denominator has a closed-form solution. For other problems, the product $L(\mathbf{x}|\theta, \Sigma)\pi(\theta, \Sigma)$ can be recognized as the kernel of the density of a well-known probability distribution. For most problems, however, the integral in the denominator lacks an analytical solution and is intractable numerically. Therefore we in most cases can only assume that we can evaluate the posterior density *up to an unknown multiplicative constant*. This common situation gave rise to the concept of *Markov Chain Monte Carlo*. Markov Chain Monte Carlo uses evaluation of the posterior density up to a multiplicative constant to create an irreducible Markov chain whose stationary distribution matches the posterior distribution. For conceptual understanding, a description of the Metropolis algorithm is given

3.2 Markov Chain Monte Carlo Algorithms

3.2.1 The Metropolis Algorithm

Recall that the objective is to create a Markov chain whose stationary distribution is the posterior distribution of a set of parameters given some data, using only evaluation of a function proportional to this posterior density $\pi^*(\theta, \Sigma|\mathbf{x}) = k\pi(\theta, \Sigma|\mathbf{x})$.

Metropolis Algorithm

Input: A function $\pi^*(\theta|\mathbf{x})$ proportional to the posterior density $\pi(\theta|\mathbf{x})$ as well as a *proposal function* P that gives a new sample θ' given θ. The proposal function is non-deterministic and is associated with a probability distribution q which must satisfy *symmetry*, that is $q(\theta'|\theta) = q(\theta|\theta')$

Output: A Markov chain with stationary distribution π

1. Initialise a random sample $\theta \in \Theta$
2. Propose a new sample $\theta' = P(\theta)$
3. Compute the *acceptance ratio* $\alpha := \frac{\pi(\theta'|\mathbf{x})}{\pi(\theta|\mathbf{x})}$. We do not know π, but $\frac{\pi(\theta'|\mathbf{x})}{\pi(\theta|\mathbf{x})} = \frac{k\pi(\theta'|\mathbf{x})}{k\pi(\theta|\mathbf{x})} = \frac{\pi^*(\theta'|\mathbf{x})}{\pi^*(\theta|\mathbf{x})}$
4. Draw $u \sim \text{Uniform}(0, 1)$. If $u < \alpha$ *accept* (append to the realisation of the Markov chain) θ'.
5. Return to 2.

From inspecting the description above one can arrive at the following intuitive understanding: A proposal that improves the posterior density is always accepted, and the probability of a proposal being rejected is proportional to how much "worse" (how much more incompatible with the prior and data) it is. One very simple (but naïve) way to define the proposal distribution is to simply perform a random walk across the parameter space. This gives Random-Walk Metropolis, which can work well for isotropic low-dimensional problems, but is known to struggle in high dimensions or for strongly anisotropic posterior distributions. This limitation is very intuitive, if there is a very small region of "correct" moves (moves to samples that do not have significantly lower posterior probability) then it is very unlikely for a random walk proposal to move in the "correct" direction along the axes where the probability increases while also not moving along axes where probability decreases. One way to resolve this problem is to decrease the variance of the random walk that generates the proposals, but this comes at the cost of increasing the autocorrelation of the chain, which is suboptimal since we for Bayesian inference require independent samples from the posterior distribution. A visualisation of this problem is given in figure. 3.1

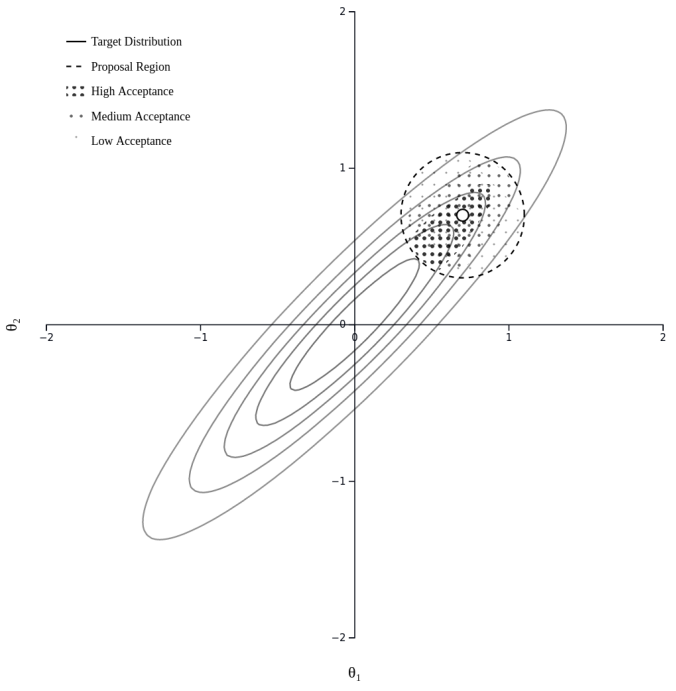

Figure 3.1: Sampling from an anisotropic distribution

To avoid the trade-off of low acceptance rate or high autocorrelation one can employ information about the gradient of the posterior distribution in order to generate proposals that conform better to it. One such method is the Metropolis-adjusted Langevin Algorithm

3.2.2 The Metropolis-adjusted Langevin Algorithm

In this section the (Euclidean) Metropolis-adjusted Langevin Algorithm[5] (henceforth MALA) will be described, as well as a possible extension of it to differentiable manifolds. We begin by defining the *Wiener process*, which is a continous-time stochastic process used in MALA

Definition 3.2.1. (Wiener process [6]) Let W_t be a continuos time stochastic process on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then W_t is said to be a *Wiener process* if the following holds

1. $W_0 = 0$ $\mathbb{P}$-almost surely.
2. If $s < t$ then $W_{t+u} - W_t \sim \mathcal{N}(0, u)$ and is independent of W_s for all $u \geq 0$.
3. $f(t) = W_t(\omega)$ is a continuos function for $\mathbb{P}$-almost every $\omega \in \Omega$.

Let $\pi^* : \mathbb{R}^d \to \mathbb{R}$ be a differentiable kernel of some posterior density function and consider the diffusion process defined by the stochastic differential equation

$$dX_t = \frac{1}{2}\nabla \log \pi^*(X) + dW_t,$$

where dW_t is the time derivative of a Wiener process W_t. It can be shown that as $t \to \infty$ the distribution of the process X_t converges to π. To avoid excessive formalism on stochastic differential equations we will avoid focusing to much on the continuos time differential equation and instead consider a discretisation of it by defining a new discrete time stochastic process $\{X_k\}_{k=1}^{\infty}$ by

$$X_{k+1} := X_k + \frac{1}{2}\tau \nabla \log \pi^*(X_k) + \sqrt{\tau}\varepsilon_k,$$

where $\varepsilon_k \sim \mathcal{N}(\mathbf{0}, I)$. This discretisation gives us the unadjusted Langevin algorithm, which unfortunately does not converge to π, although the introduced discretisation error is linearly bounded[1]. The discretisation error can be removed by introducing a Metropolis-Hastings step where the unadjusted Langevin process generates the proposals. This gives us the Metropolis-adjusted Langevin Algorithm in $\mathbb{R}^d$. We will now briefly discuss the Hastings correction in the acceptance ratio for Euclidean MALA, in order to motivate how the acceptance ratio for the Riemannian case is derived.

Let x_k denote the realisation at step k of the Markov chain given by the Metropolis Adjusted Langevin Algorithm. Clearly, we have that

$$X_{k+1} \sim \mathcal{N}\left(x_k + \frac{1}{2}\tau \nabla \log \pi^*(x_k), \tau I\right).$$

This leads naturally to the insight that the kernel of the proposal density is given by

$$q(x_{k+1}|x_k) \propto e^{-\frac{1}{2\tau}||x_{k+1} - x_k - \frac{1}{2}\nabla \log \pi^*(x_k)||_2^2}.$$

For Euclidean space, the metric tensor is global (independent of the point at which one is evaluating it) and given in matrix form by the identity matrix. For the Riemannian case we will however see that the covariance matrix of the proposal will instead become dependent on the current sample since it will be the inverse of the metric tensor at the current point.

[1] Formally, one can by the triangle inequality bound the Wasserstein distance between the target distribution and asymptotic sampling distribution linearly in the step size τ

3.2.3 Generalisation to Riemannian Manifolds

There are several ways to generalise the discretisation of Langevin diffusion to a differentiable manifold. We will consider the following generalisation:

Definition 3.2.2. (Discretised Langevin Diffusion on a Differentiable Manifold). Let $E_1, E_2 \dots E_n$ denote basis vectors of some orthonormal basis of $T_X\mathcal{M}$ and let $\zeta \sim \mathcal{N}(\mathbf{0}, I)$. Furthermore let τ be the *step size*. Then

$$X_{k+1} = \exp_X \left(\tau \nabla_{\mathcal{M}} \log \pi(X_k) + \sqrt{2\tau} \zeta^i E_i \right)$$

defines discretised Langevin diffusion on $\mathcal{M}$[2] ($\zeta^i E_i$ is Einstein summation notation) [3]. Any finite realisation of the process defined above is an unadjusted Langevin algorithm on $\mathcal{M}$ [16] [3]

3.2.4 Mixing

A term that often arises in the field of Markov chain Monte Carlo is *mixing*. There are several ways to construct Markov chains whose stationary distribution equals some target distribution (in fact, for most cases, Random Walk Metropolis will have this property). However, asymptotics are never appliceable in practice by definition. We are instead interested in designing Markov chains that can be stopped in finite (ideally also reasonable) time so that we can use their samples for inference. A chain that has this property is said to be *rapid mixing*. To understand why rapid mixing is desirable and why high autocorrelation causes problems when we wish to perform inference we will state the Markov Chain Central Limit Theorem

Theorem 3.2.3. *(Markov Chain Central Limit Theorem) [2] Suppose we wish to estimate* $\mathbb{E}(g(X_1))$ *for some measurable function* g *where* $X_1 \sim \mathcal{D}$. *We have samples* $X_1, X_2, X_3 \cdots X_n$ *from a Markov Chain whose stationary distribution is* $\mathcal{D}$. *Let* $\mu = \mathbb{E}(g(X_1))$ *and let* $\hat{\mu}_n := \frac{1}{n}\sum_{k=1}^{n} g(X_k)$. *Then*

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2),$$

where

$$\sigma^2 = \mathrm{Var}(g(X_1)) + 2\sum_{k=1}^{\infty} \mathrm{Cov}(g(X_1), g(X_{k+1})).$$

From the condition that $X_1 \sim \mathcal{D}$ we see that *rapid mixing* is required, we want the Markov chain to approximately reach its stationary distribution in short time so that subsequent samples become "useable" for estimation (that is, they roughly satisfy the conditions in Theorem 3.2.3). Furthermore, the faster the autocorrelation decays the lower the variance becomes, meaning fewer samples are required for accurate estimation of μ. As previously stated the mixing properties of MALA and HMC are generally superior to those of RWM. For the generalisation of MALA to Riemannian manifolds, the discretisation error impacts the mixing properties of the resulting Markov chain. In section 4 we prove that the manifold of symmetric positive definite matrices has certain properties (defined in [3]) that guarantee this discretisation error is bounded. More specifically, it is shown that the *iteration complexity* (number of iterations of the RMALA algorithm) grows in $\mathcal{O}(\epsilon^{-2})$ where ϵ is the desired *Wasserstein distance* between the invariant measure (stationary distribution of the continous-time diffusion) and the distribution of the samples after some amount of iterations. The Wasserstein distance is defined in the following way:

Definition 3.2.4. (Wasserstein distance) Let μ, ν be two probability measures on a metric space (M, d) [4]. The Wasserstein distance between μ and ν is given by

$$W(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \mathbb{E}_\gamma[d(x, y)]$$

where Γ is the set of all *couplings* of μ, ν. A coupling of μ, ν is a joint distribution $M \times M$ where the marginal distributions of the components are μ, ν.

[2] The proper way to sample a diffusion process uses the Laplace-Beltrami operator, but this yields a decent approximation for small τ

[3] This can intuitively be thought of as "Diffuse in the tangent space and project back".

[4] we require the metric space to be a Polish space, but all Riemannian manifolds are trivially Polish spaces so we avoid this formal technicality in the interest of brevity

The Wasserstein distance has the properties of a *metric*[5]: $W(\mu,\nu) = 0$ if and only if $\mu = \nu$, $W(\mu,\nu) > 0$ for all $\mu \neq \nu$, $W(\mu,\nu) = W(\nu,\mu)$ for all μ,ν and $W(\mu,\nu) \leq W(\mu,\nu') + W(\nu',\nu)$ (the last condition is the triangle inequality). We take a particular interest in the Wasserstein distance due to the following fact:

Theorem 3.2.5. *(Convergence in Wasserstein distance is stronger than convergence in distribution, [15]). Let $\{X_n\}_{n=1}^{\infty}$ be a sequence random variables and denote by μ_n the distribution of X_n. Furthermore let Y be a random variable with distribution given by μ. Then*

$$W(\mu_n,\mu) \to 0 \text{ as } n \to \infty \Rightarrow X_n \xrightarrow{d} Y \text{ as } n \to \infty$$

Theorem 3.2.5 makes it immediately clear why we are interested in bounding the Wasserstein distance: Informally, if we can bound the Wasserstein distance above between the distribution of our samples and our target distribution by some small ϵ, then these samples will approximately have the same distribution as our target distribution, meaning that we can appeal to the Markov Chain Central Limit Theorem (3.2.3) to motivate estimating measurable functions of random variables with our target distribution by evaluating them on our MCMC samples. For our purposes, if we can show that conditions I - IV in [3] hold, we obtain guarantees on the order of the iteration complexity required for a specific Wasserstein distance bound for our MALA samples.

[5]As noted in [15] this is technically not true, as the Wasserstein distance between two measures can be infinite, so the "metric" is not real-valued. This does not change any practical conclusions, however.

Chapter 4

Theoretical Mixing Results

> The theory of probabilities is basically just common sense reduced to calculus.
>
> –Pierre-Simon Laplace

In this section we will investigate four sufficient conditions for theoretical guarantees on mixing time and convergence of the Metropolis Adjusted Langevin Algorithm on Riemannian manifolds. In the paper "Efficient Sampling on Riemannian Manifolds via Langevin MCMC" [3] two conditions relating to the geometry of the manifold and two conditions relating to the potential function of the target distribution are given. If these conditions are satisfied it is shown that the Wasserstein distance between the distribution of the samples and the target distribution can be bounded above by ϵ after $\mathcal{O}(\epsilon^{-2})$ steps. Before introducing these, we make a short note on common generalisations of the Lewandowski-Kurowicka-Joe distribution (henceforth LKJ distribution)

4.1 Generalisations of the LKJ distribution

The LKJ distribution is a common prior distribution to use over the space of **correlation** matrices, and is defined as follows

Definition 4.1.1. (LKJ distribution) We say that $\mathcal{C} \sim \text{LKJ}(\eta)$ if the density for a realisation C is given by

$$p(C) \propto (\det C)^{\eta-1}$$

where $\eta \in (0, \infty)$ is the shape parameter[1].

In this thesis (and in many practical applications in Bayesian modelling) we are interested in estimating the covariance matrix, not the correlation matrix. Any covariance matrix can be decomposed into a diagonal matrix containing the marginal variances and a correlation matrix in the following way:

$$\Sigma = \underbrace{\begin{pmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{nn} \end{pmatrix}}_{D} \underbrace{\begin{pmatrix} 1 & r_{12} & \cdots & r_{1n} \\ r_{21} & 1 & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & 1 \end{pmatrix}}_{C} \underbrace{\begin{pmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{nn} \end{pmatrix}}_{D}.$$

One can then obtain a distribution for Σ by considering it as a joint distribution over the components D, C where the LKJ distribution is the marginal distribution for C, with some freedom to choose the marginal over the components of D. One can, for example choose D to be the χ-distribution, which is strongly log-concave.

[1] For some intuition regarding the shape parameter, $\eta = 1$ gives a uniform distribution over correlation matrices, $\eta > 1$ places more mass on weak correlations and $\eta < 1$ places more mass on strong correlations.

4.2 Sufficient conditions for bounding the iteration complexity

The sufficient conditions given in [3] for bounding the iteration complexity of RMALA with respect to some Wasserstein bound ϵ by $\mathcal{O}(\epsilon^{-2})$ are as follows

1. We assume that for all $x \in \mathcal{M}$ and all tangent vectors $u \in T_x\mathcal{M}$ there exists a global constant $L_{Ric} \in \mathbb{R}$ such that $\text{Ric}(u,u) \geq -L_{Ric}||u||^2$[2]

2. A vector field β is said to be $(m, q, \mathcal{R})$-distant-dissipative if there exists real $m, q, \mathcal{R}$ with $m > 0$ and $\mathcal{R} \geq 0$ such that for all x, y satisfying $d(x,y) \geq \mathcal{R}$ there exists a minimising geodesic $\gamma : [0,1] \to M$ such that $\gamma(0) = x$ and $\gamma(1) = y$ such that the inequality $\langle \Gamma(\beta(y); y \to x) - \beta(x), \gamma'(0)\rangle \leq -md(x,y)^2$. Here β is the vector field defined as $-\frac{1}{2}\nabla h$ where h is the negative log-density of the distribution we are sampling from. Γ denotes parallel transport from T_yM to T_xM along γ. It is also assumed that for all $x, y \in \mathcal{M}$ satisfying $d(x,y) \leq \mathcal{R}$ it holds that $\langle \Gamma_{yx}\beta(y) - \beta(x), \gamma'(0)\rangle \leq qd(x,y)^2$

3. A vector field β is said to be L'_β-Lipschitz if we have for all$x \in \mathcal{M}$ and all $u \in T_x\mathcal{M}$ $||\nabla_v \beta(x)|| \leq L'_\beta||u||$

4. Let R be the Riemannian curvature tensor of $\mathcal{M}$. We assume there exists some $L_R \in \mathbb{R}^+$ such that for all $x \in \mathcal{M}$, and for all $u, v, w, z \in T_x\mathcal{M}$ we have that $\langle R(u,v)v, u\rangle \leq L_R||u||^2||v||^2$ [3]

We will focus our discussion on condition I, the first half of condition II and condition IV. This is due to the second half of condition II being redundant for strictly convex potentials, and since not a lot of theory can be developed from a simple Lipschitz condition, which is condition III.

4.2.1 Condition I

We begin by proving statement 1. First, let $\mathcal{M} := M_1 \times M_2$ and denote by $\pi_i : \mathcal{M} \to M_i$ the projection from the product manifold $\mathcal{M}$ to the factor manifold M_i.

Theorem 4.2.1. *(The metric tensor for a product manifold is a block tensor, see [9]) Let (M_1, g_1) and (M_2, g_2) be two Riemannian manifolds and define the product manifold $(M_1 \times M_2, g)$. Then the Riemannian metric on the product manifold is given by the tensor sum $g_1 \oplus g_2$*

Theorem 4.2.2. *Let $\text{Ric}_1(u,v)$ be the Ricci curvature tensor on M_1 and let $\text{Ric}_2(u,v)$ be the Ricci curvature tensor on M_2 The Ricci curvature tensor on the product manifold $M = M_1 \times M_2$ will simply be the tensor sum of the curvature tensors on the factor manifolds. Since these are $(0,2)$-tensors this means in particular that $(\text{Ric}_1 \oplus \text{Ric}_2)(u,v) = \text{Ric}_1(\pi_1 u, \pi_1 v) + \text{Ric}_2(\pi_2 u, \pi_2 v)$ [7]*

Lemma 4.2.3. *The Ricci curvature of the manifold $M = \mathbb{R}^k \times \mathcal{P}(k)$ can be computed as $\text{Ric}_M(u,v) = \text{Ric}_{\mathcal{P}(k)}(\pi u, \pi v)$ where π denotes the projection of M onto $\mathcal{P}(k)$*

Proof. Follows immediately from the fact that $\mathbb{R}^k$ has zero curvature everywhere and theorem 4.2.2 □

Theorem 4.2.4. *(Explicit construction of the Ricci curvature tensor [13]). Consider the basis for the tangent space of $\mathcal{P}(n)$ defined in [13]. In this basis the Ricci curvature tensor can be represented in the following way*

$$\text{Ric}_{\mathcal{P}(k)}(u,v) = -\frac{n}{4}u^t \begin{pmatrix} I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T & 0 \\ 0 & I_{n(n-1)/2} \end{pmatrix} v.$$

We are now ready to prove our first important result

[2]The $||u||^2$ term does not appear in the original paper. E-mail correspondence with one of the authors (Dr. Cheng) confirms this is an erroneous omission and that the $||u||^2$ term is in fact necessary

Theorem 4.2.5. *(M satisfies condition 1) Let $M = \mathbb{R}^k \times \mathcal{P}(k)$. There exists some scalar L_{ric} such that $\forall x \in M, u \in T_x M$ it holds that* $\text{Ric}(u,u) \geq -L_{Ric}$

Proof. By lemma 4.2.3 we have that $\text{Ric}(u,u) = \text{Ric}_{\mathcal{P}(k)}(\pi u, \pi u)$ which by theorem 4.2.4 can be represented as $(\pi u)^T A(\pi u)$ where

$$A = -\frac{n}{4}\begin{pmatrix} I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T & 0 \\ 0 & I_{n(n-1)/2} \end{pmatrix}$$

for a certain choice of basis on $T_x M$. Since A does not depend on x we can obtain a uniform lower bound on the quadratic form, expressed in the eigenvalues of A □

4.2.2 Condition II

We now move on to condition II. We will begin by showing that μ-strong geodesic convexity is a stronger condition than $(m, q, \mathcal{R})$-distance dissipativity. It will also be shown that geodesically linear functions can never be $(m, q, \mathcal{R})$-distance dissipative, and that the Log-Sobolev inequality always fails for weakly convex vector fields on manifolds with a non-positive lower bound on their Ricci curvature. This allows us to exclude a large class of matrix variate distributions from attaining the mixing rate guarantees presented in [3]

Geodesic convexity

We begin by defining μ-strong geodesic convexity

Definition 4.2.6. Let $f : \mathcal{M} \to \mathbb{R}$. Then f is said to be geodesically μ-strongly convex if it for any $x, y \in \mathcal{M}$ holds that

$$f(y) \geq f(x) + \langle \nabla f(x), \text{Log}_x(y) \rangle_x + \frac{\mu}{2} d(x,y)^2.$$

here Log denotes the logarithmic map, that is the inverse of the exponential map. [17]

Lemma 4.2.7. *(Gradient-Log inequality) Suppose that f is μ-strongly geodesically convex on $\mathcal{M}$. Then for any two points $x, y \in \mathcal{M}$ it holds that*

$$\langle \nabla f(y), \text{Log}_y(x) \rangle_y + \langle \nabla f(x), \text{Log}_x(y) \rangle_x \leq -\mu d(x,y)^2. \tag{4.1}$$

Proof. We have by assumption the two inequalities

$$f(y) \geq f(x) + \langle \nabla f(x), \text{Log}_x(y) \rangle_x + \frac{\mu}{2} d(x,y)^2$$
$$f(x) \geq f(y) + \langle \nabla f(y), \text{Log}_y(x) \rangle_y + \frac{\mu}{2} d(x,y)^2$$

If we in the first inequality replace $f(x)$ with the right hand side in the second inequality we obtain

$$f(y) \geq \underbrace{f(y) + \langle \nabla f(y), \text{Log}_y(x) \rangle_y + \frac{\mu}{2} d(x,y)^2}_{\text{Right hand side of second inequality}} + \langle \nabla f(x), \text{Log}_x(y) \rangle_x + \frac{\mu}{2} d(x,y)^2$$

Which after simplification gives

$$\langle \nabla f(y), \text{Log}_y(x) \rangle_y + \langle \nabla f(x), \text{Log}_x(y) \rangle_x \leq -\mu d(x,y)^2.$$

□

We are now ready to prove that μ-strong convexity implies condition two holds for all $\mathcal{R} \geq 0$

Theorem 4.2.8. *On a symmetric space $\mathcal{M}$, μ-strong convexity is a stronger condition than $(m, q, \mathcal{R})$-distance dissipativity.*

Proof. Recall that if $\Gamma_{yx} : T_y\mathcal{M} \to T_x\mathcal{M}$ is a parallel transport map then Γ_{yx} is not only a linear isomorphism but also an isometry of the inner product spaces $T_y\mathcal{M}$ and $T_x\mathcal{M}$. This means we can convert the inner product in the first term of equation (4.1) given in lemma 4.2.7 to an inner product on the tangent space $T_x\mathcal{M}$. This gives us

$$\langle \Gamma_{yx}\nabla f(y), \Gamma_{yx}\operatorname{Log}_y(x)\rangle_x + \langle \nabla f(x), \operatorname{Log}_x(y)\rangle_x \leq -\mu d(x,y)^2.$$

Now, by assumption $\mathcal{M}$ is a *symmetric space* (See section 6.5 in [1]) meaning that $\Gamma_{yx}\operatorname{Log}_y(x) = -\operatorname{Log}_x(y)$ giving us

$$\langle \nabla f(x), \operatorname{Log}_x(y)\rangle_x - \langle \Gamma_{yx}\nabla f(y), \operatorname{Log}_x(y)\rangle_x \leq -\mu d(x,y)^2.$$

Applying the linearity of real inner products we obtain

$$\langle \nabla f(x) - \Gamma_{yx}\nabla f(y), \operatorname{Log}_x(y)\rangle_x \leq -\mu d(x,y)^2.$$

Now using that $\operatorname{Log}_x(y) = \gamma'(0)$ where γ is a minimising geodesic connecting x and y we obtain

$$\langle \nabla f(x) - \Gamma_{yx}\nabla f(y), \gamma'(0)\rangle_x \leq -\mu d(x,y)^2.$$

Finally since the vector field defined as the negative of the gradient we obtain that condition 2 is implied by μ-strong convexity. □

Geodesic linearity

We will now characterise the geodesically linear functions on $\mathcal{P}(n)$

Definition 4.2.9. (Geodesic linearity) A function $f : \mathcal{M} \to \mathbb{R}$ is said to be geodesically linear if the following holds

$$(f \circ \gamma)(t) = at + b$$

where γ is any geodesic and a, b are real constants.

A very important result for the continued analysis of the mixing properties of Langevin diffusion for potentials induced by matrix-variate prior distributions is the following:

Lemma 4.2.10. *(Geodesic linearity of the log-determinant) The function* $f(\Sigma) := \log\det\Sigma$ *is geodesically linear*

Proof. Let $\gamma : [0,1] \to \mathcal{M}$ be a geodesic connection Σ_1, Σ_2. We have

$$\gamma(t) = \Sigma_1^{1/2}(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2})^t\Sigma_1^{1/2}$$

Using the additive properties of the log-determinant function we get

$$\log\det\gamma(t) = \log\det\Sigma_1 + t\log\det\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}$$

which satisfies definition 4.2.9 with $a = \log\det\Sigma_1$ and $b = \log\det\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}$ □

We are now ready to prove the important result that a large class of matrix variate distributions are not suitable as potential functions for Riemannian Langevin diffusion

Theorem 4.2.11. *(Impossibility of* $(m, q, \mathcal{R})$*-distance dissipativity for induced potentials of geodesically linear functions). Let* $f : \mathcal{M} \to \mathbb{R}$ *be geodesically linear. Then there exists no* $m, q, \mathcal{R}$ *such that*

$$\langle \Gamma_{yx}\nabla f(y) - \nabla f(x), \gamma'(0)\rangle_x \leq -md(x,y)^2.$$

Proof. In the interest of brevity we give the proof idea, and allow the interested reader to fill in the details. The argument is as follows: Since f is geodesically linear its Riemannian gradient will be a Killing vector field[3]. The parallel transport map restricted to a Killing field is the identity map, meaning that the term $\Gamma_{yx}\nabla f(y) - \nabla f(x) = 0$ and thus the inner product vanishes everywhere, meaning we can never upper bound it by a negative definite expression in x, y □

[3]One might think it is called a Killing vector field since it preserves isometries, thus killing off any curvature terms, but it gets its name from the mathematician Wilhelm Killing.

This immediately gives us that if the marginal distribution over correlation matrices is the LKJ distribution we will fail to satisfy the conditions for mixing guarantees. With the additional fact that the trace operator is only weakly geodesically convex on $\mathcal{P}(n)$ we conclude that the Wishart distribution also fails to satisfy these conditions globally (that is, for $\mathcal{R} = 0$). The Riemannian Gaussian distribution, defined as

$$p(\Sigma) \propto e^{-\tau d(\Sigma,\Sigma_0)^2}$$

will however satisfy this condition since by lemma 4.8.2 in [8] the Riemannian metric on a manifold with non-positive sectional curvature is 1-strongly geodesically convex.

4.2.3 Condition IV

Lemma 4.2.12. *(Relationship of sectional curvature and Riemannian curvature) Let $\mathcal{M}$ be a Riemannian manifold. If the sectional curvature of $\mathcal{M}$ is everywhere non-positive, the quadratic form $(u, v) \mapsto \langle R(u, v)v, u\rangle$ is negative semidefinite*

Proof. Recall that the sectional curvature of linearly independent tangent vectors is given by

$$K(u, v) = \frac{\langle R(u, v)v, u\rangle}{\langle u, u\rangle\langle v, v\rangle - \langle u, v\rangle^2}$$

By the Cauchy-Schwartz inequality the denominator is always positive, and thus if $K(u, v)$ is negative then $\langle R(u, v), u\rangle$ must be negative. For linearly dependent tangent vectors the skew-symmetry of $R(u, v)$ gives that $R(u, \lambda u) = 0$ and therefore $K(u, v) < 0$ implies negative semi-definiteness of the aforementioend quadratic form induced by the Riemannian curvature tensor □

Theorem 4.2.13. *($\mathbb{R}^n \times \mathcal{P}(n)$ satisfies condition IV)*

Proof. Follows immediately from the fact that the Riemannian curvature tensor on the product manifold is the direct sum of the curvature tensors on the constituent manifolds and that $\mathcal{P}(n)$ is a Cartan-Hadamard manifold □

We are now ready to state the central result of this thesis

Theorem 4.2.14. *(Mixing time of the Riemannian Metropolis Adjusted Langevin Algorithm on $\mathbb{R}^n \times \mathcal{P}(n)$) Let $\pi : \mathbb{R}^n \times \mathcal{P}(n) \to \mathbb{R}$ be defined as $\pi(x) \propto e^{-\beta(x)}$ for some geodesically μ-strongly convex potential function β. Then the Wasserstein distance between the distribution of samples from the Riemannian Metropolis Adjusted Langevin Algorithm and the invariant measure $\pi(x)$ will be bounded above by ϵ after $\mathcal{O}(\epsilon^{-2})$ iterations*

Proof. Follows from theorems 4.2.5, 4.2.8 and 4.2.13 as well as the general result given in [3] □

To shed light on this central theoretical result, we will refer back to the Wishart, χ-LKJ and Riemannian Gaussian distributions (see Table 4.1)

$\propto p(\Sigma)$	**Name**	**Mixing bound**
$\lvert\Sigma\rvert^{(n-p-1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\Sigma_0^{-1}\Sigma)\right)$	Wishart density	Fails the stronger condition of global μ-strong geodesic convexity
$\prod_{i=1}^p d_i^{\eta} \lvert R\rvert^{-1} \prod_{i=1}^p \sigma_i^{p-1}$	χ-LKJ density	Can never satisfy II, even outside of a compact subset due to geodesic linearity of $\log\det$
$\exp(-\frac{1}{\sigma_0^2} d^2(\Sigma, \Sigma_0))$	Riemannian Gaussian density	By the strong convexity of the Riemannian distance function, this satisfies condition II. By the smoothness of the distance function it satisfies condition III

Table 4.1: Properties of prior distributions supported on $\mathcal{P}(n)$

This theoretical result suggests that the most appropriate prior distribution for problems where one wishes to utilise the Riemannian Metropolis Adjusted Langevin Algorithm is the Riemannian Gaussian distribution. It should be noted, however, that the conditions given in [3] are sufficient, but not necessarily necessary. As a concluding remark to this thesis, we therefore conjecture the following

Conjecture 4.2.15. Let $\mathcal{D}_{\cdot|\Sigma_0}$ denote the conditional distribution of the sample Σ_k from discretised Langevin diffusion process on $\mathcal{P}(n)$, where the conditioning is with respect to the initial sample Σ_0. If

$$\langle \Gamma_{\Sigma_k \Sigma_0} \beta(\Sigma_k) - \beta(\Sigma_0), \gamma'(0) \rangle \leq -m d(\Sigma_k, \Sigma_0)^2$$

holds $\mathcal{D}_{\cdot|\Sigma_0}$-almost surely for some positive constant m and some $\Sigma_0 \in \mathcal{P}(n)$ then (provided conditions I, III and IV hold) one obtains the same $\mathcal{O}(\epsilon^{-2})$ bound on the iteration complexity of RMALA.

The intuition for this conjecture is as follows: condition II is required to ensure the integrability of $e^{-\beta(x)}$, that is to prevent the diffusion process from indefinitely drifting across the manifold, never converging to a compact subset. Condition II essentially requires that the potential induces contractive behaviour on the diffusion process to prevent this indefinite drift. But intuitively, if the circumstances which cause this bound to fail happen with probability 0 under the diffusion process it should not in practice effect this integrability, since we will almost always have contraction toward a region of high probability. Furthermore, it is shown in [3] that one obtains the same bounds even when replacing the exact gradient with a stochastic estimate of it[4], which should further heuristically convince the reader that weakening the conditions to hold only probabilistically would not make a practical difference. Such a re-formulation would likely be a fruitful future endeavour.

[4] akin to stochastic gradient descent in machine learning

Bibliography

[1] Rajendra Bhatia. *Positive Definite Matrices*. Princeton Series in Applied Mathematics. Princeton University Press, 2007.

[2] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng, editors. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 1 edition, 2011.

[3] Xiang Cheng, Jingzhao Zhang, and Suvrit Sra. Efficient sampling on riemannian manifolds via langevin mcmc, 2024.

[4] Chad M. Eliason, Lauren E. Mellenthin, Taylor Hains, Jenna M. McCullough, Stacy Pirro, Michael J. Andersen, and Shannon J. Hackett. Genomic signatures of convergent shifts to plunge-diving behavior in birds. *Communications Biology*, 6(1):1011, October 2023.

[5] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(2):123–214, 03 2011.

[6] Geoffrey R Grimmett and David R Stirzaker. *Probability and Random Processes*. Oxford University Press, 05 2001.

[7] Andrew D. Hwang. Ricci curvature on product manifold. Mathematics Stack Exchange, May 2021. URL: https://math.stackexchange.com/questions/4149567/ricci-curvature-on-product-manifold.

[8] Jürgen Jost. *Riemannian Geometry and Geometric Analysis*. Universitext. Springer, 5 edition, 2008.

[9] John M. Lee. *Riemannian Manifolds: An Introduction to Curvature*, volume 176 of *Graduate Texts in Mathematics*. Springer, 1997.

[10] Mouagip. Aminoacids table.svg. Wikimedia Commons, February 2009. Public domain image.

[11] Ben Murrell, Sasha Moola, Amandla Mabona, Thomas Weighill, Daniel Sheward, Sergei L. Kosakovsky Pond, and Konrad Scheffler. Fubar: A fast, unconstrained bayesian approximation for inferring selection. *Molecular Biology and Evolution*, 30(5):1196–1205, 02 2013.

[12] S V Muse and B S Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11(5):715–724, Sep 1994.

[13] Xavier Pennec. 3 - manifold-valued image processing with spd matrices. In Xavier Pennec, Stefan Sommer, and Tom Fletcher, editors, *Riemannian Geometric Statistics in Medical Image Analysis*, pages 75–134. Academic Press, 2020.

[14] Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006.

[15] Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin, 2009.

[16] Xiao Wang, Qi Lei, and Ioannis Panageas. Fast convergence of langevin dynamics on manifold: Geodesics meet log-sobolev. *arXiv preprint*, 2020.

[17] Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *JMLR: Workshop and Conference Proceedings*, volume 49, pages 1–21. JMLR, 2016.