The Riemannian Langevin Equation and Its Applications in Random Matrix Theory and Gibbs Sampling Problems

By

TIANMIN YU

B.S., Peking University, 2018

B.Econ, Peking University, 2018

Thesis

Submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in the Department of Applied Mathematics at Brown University

PROVIDENCE, RHODE ISLAND

May 2024

This dissertation by Tianmin Yu is accepted in its present form by the Division of Applied Mathematics as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date____________________ ____

Govind Menon, Ph.D., Advisor

Recommended to the Graduate Council

Date____________________ ____

Jerome Darbon, Ph.D., Reader

Date____________________ ____

Yair Shenfeld, Ph.D., Reader

Approved by the Graduate Council

Date____________________ ____

Thomas A. Lewis, Dean of the Graduate School

Acknowledgements

I would like to express my deepest gratitude to my committee chair Prof. Govind Menon for his continuous support on every side of my academic life. It is my fortunate to work with you focusing on problems that are really interesting to me. I also want to thank Prof. Jerome Darbon and Prof. Yair Shenfeld for agreeing to read my thesis and providing precious comments. I also had great pleasure of working with Prof. Xiangxiong Zhang and Prof. Jianfeng Lu, thank you for your guidance in our collaboration and your appreciation. Thanks should also go to my colleagues Tejas Kotwal, Jake Mundo and Zsolt Veraszto. I appreciate the time we talk freely about every aspect of our research.

I'm deeply indebted to my family and friends. I thank my parents for trusting me to pursue my life. Special thanks go to my fiancee Peiwen Zhu for her companion and spiritual support all the time. I thank my friends Dr. Xiaoyu Xie, Dr. Ziyao Xu, Dr. Enrui Zhang, Dr. Zhen Zhang, Mengjie Liu, Dr. Tong Shen, Dr. Kun meng, Qian Zhang, Sicheng Liu, Xin Lian and Xiaoding Yang for their help throughout the past 5 years. In particular, I really enjoy the time with Sicheng Liu and Xiaoding Yang chatting in our office about everything mathematical or non-mathematical. Thank all my friends for the time we spend in Providence.

Contents

0 Overview **1**

1 The Riemannian Langevin Equation **5**

1.1 Background 6

1.2 SDE on manifolds 9

1.3 Transition properties of Brownian motion 15

1.4 Riemannian Langevin Equation 22

2 Siegel Brownian motion **25**

2.0 Abstract 25

2.1 Introduction 27

2.1.1 Outline 27

2.1.2 Motion by curvature and stochastic flows 29

2.1.3 General principles 30

2.1.4 A new matrix model 31

2.2 Background: SDE and martingales on manifolds 36

2.2.1 Outline 36

2.2.2 SDE, martingales on manifolds 36

2.2.3 Brownian motion and orthogonal projections 38

2.2.4 Some matrix algebra to be used 41

2.3 Proof of Theorem 3 45
2.3.1 The unit disk $\mathcal{D}$ 45
2.3.2 The orthonormal basis 47
2.3.3 Covariant derivatives 51
2.3.4 Evolution of spectrum 53
2.4 Proof of Theorem 4 56
2.4.1 Orbit volume 56
2.4.2 Metric and gradient flow 59
2.5 Well-posedness and Non-colliding property 60

3 Gibbs sampling problem on PSD matrices **65**
3.0 Abstract 65
3.1 Introduction 66
3.1.1 Problem statement 66
3.1.2 Two special Riemannian metrics of $S_+^{n,p}$ 68
3.1.3 Riemannian Langevin equation 68
3.1.4 Connection to existing works and contributions 70
3.2 Brownian motion constructed by smooth orthonormal frames 72
3.2.1 Notation 72
3.2.2 Brownian motion as a solution to SDE 74
3.2.3 Brownian motion on an embedded submanifold 77
3.2.4 Brownian motion via Riemannian submersion 80
3.3 Explicit Riemannian Langevin equation on $\mathcal{S}_+^{n,p}$ 84
3.3.1 Embedded metric case 84
3.3.2 Bures-Wasserstein metric case 88
3.4 Convergence analysis of RLE 95
3.4.1 Convergence to Gibbs 95
3.4.2 Rate convergence: review of general results 96

3.4.3 Bakry-Émery-Ricci tensor for the embedded geometry 100
3.4.4 Bakry-Émery-Ricci tensor for the Bures-Wasserstein metric . . 106
3.4.5 Numerical illustration 108

4 Numerical schemes for Gibbs sampling problem 110

4.0 Abstract 110
4.1 Introduction 111
4.1.1 Problem statement 111
4.1.2 Two Riemannian metrics on $\mathcal{S}_+^{n,p}$ 112
4.1.3 Langevin dynamics and the Riemannian Langevin equation . . 113
4.1.4 Riemannian Langevin Monte Carlo sampling schemes 115
4.1.5 Some applications and related work 117
4.1.6 Organization of the paper 119
4.2 Riemannian Langevin equations on $\mathcal{S}_+^{n,p}$ 120
4.2.1 The Riemannian Langevin equation for embedded geometry $(\mathcal{S}_+^{n,p}, g_E)$ 120
4.2.2 The Riemannian Langevin equation for Bures-Wasserstein geometry $(\mathcal{S}_+^{n,p}, g_{BW})$ 122
4.3 Two Riemannian Langevin Monte Carlo schemes 125
4.3.1 Scheme E for the embedded geometry 125
4.3.2 Scheme BW for the Bures-Wasserstein metric 129
4.4 Examples with analytical formulae 132
4.4.1 Scalar random variables 132
4.4.2 Example I: $\mathcal{E}(X) = \frac{1}{2}||X||_F^2$ 134
4.4.3 Example II: $\mathcal{E}(X) = \mathrm{Tr}(X \log X)$ 135
4.4.4 Example III: $\mathcal{E}(X) = \frac{1}{2}||X - A||_F^2$ 136
4.4.5 MCMC numerical integration 139
4.5 Numerical tests 141

4.5.1 Numerical validation of the scalar variable $D(X)$ 142
4.5.2 MCMC numerical integration 144
4.5.3 A numerical study of the convergence to equilibrium 146
4.6 Conclusion 149

5 RLE in conic programming 150
5.1 Introduction 150
5.1.1 Outline 150
5.1.2 The canonical metric 151
5.1.3 Riemannian Langevin equation 152
5.1.4 Riemannian geometries in optimization 153
5.1.5 A comparison of SGD with RLE 155
5.2 Brownian motion and conic programs 157
5.3 Brownian motion examples 160
5.3.1 Positive orthant 160
5.3.2 Cube 160
5.3.3 Lorentz cone 161
5.4 Additional properties 164
5.4.1 Divergence of volume 164
5.4.2 Mean curvature of level surface 165

6 RLE in Deep Linear Network 167
6.1 Introduction 167
6.1.1 Deep Linear Network 167
6.1.2 Balanced manifold of DLN 168
6.2 Main results 170
6.3 Geometric analysis of $\mathcal{M}$ 172
6.4 Proof of main results 181

7 Summary **185**

Chapter 0

Overview

This dissertation contains a preliminary part and several parts of applications in different problems. The first part introduces Brownian motion on manifolds and Riemannian Langevin Equation. The other parts include applications in random matrix theory, Gibbs sampling problem, Conic programming, and deep linear network. In each parts we are in the situation where there is a microscopic space and a macroscopic space, and we can use the transition property of Brownian motion to operate on the microscopic space to affect macroscopic space.

The Riemannian Langevin Equation (RLE) lies in the interdisciplinary field of probability theory, Riemannian geometry, and optimization problems. Brownian motion on Riemannian manifolds was used to solve geometric problems in probabilistic methods in the 1980s, with the advantage of being more intuitive and explicit. On the optimization side, the RLE is the natural relaxation of gradient descent, giving more consideration to the geometric properties of the underlying space rather than solely the objective function. The evolution of a point through the flow is replaced by the diffusion of a probability distribution, and the optimal solution is replaced by an equilibrium measure. Such relaxation makes a trade-off between precision, speed, and other flexibilities, e.g., less dependence on initialization, avoiding being stuck at saddle points or local minima.

The stochastic feature of the RLE also leads to phenomena that will never appear in gradient flow. Conserved quantities in gradient flow give directions of motion that do not affect the optimization process, so a reduced system with cyclic variables removed is desired. However, in the RLE, the diffusion of cyclic variables still produces mean curvature flow and will twist the target function in the reduced system. This phenomenon coincides with the minimal free energy principle (instead of energy) in thermodynamics and could be explained as the effect of entropy. In the chapter 1, we will further discuss this relationship and find that it is congruent with Boltzmann's explanation of entropy.

This thesis contains several concrete examples and applications of Brownian motions on different manifolds, which connect the theory of Riemannian Langevin equation and stochastic numerical algorithms based on it. Comparing with existing literature about RLE, the main advantage of these works is that the geometric structure of the problem is emphasized, in the sense that all motivation and intuition are coming from geometric principle, and proofs are all based on analysis of Riemannian geometry.

The organization of this thesis is as follows: The first chapter 1 introduces the main concepts and properties of Brownian motion on Riemannian manifolds and the Riemannian Langevin equation. The following chapters are joint work papers with different collaborators.

Chapter 2 is a joint work with Govind Menon on application of Brownian motion on Random matrix theory. In this paper, a geometric framework for constructing stochastic processes which generate deterministic processes is discussed and applied to Siegel upper half space, a canonical Symplectic space with hyperbolic geometry.

This framework is based on Riemannian submersion and the deterministic motion is driven by mean curvature. In the example of Siegel upper half space, the Brownian motion is explored and the evolution of spectrum gives a hyperbolic variant of Dyson Brownian motion. This work may be interesting to people working on Symplectic geometry and probability theory.

Chapter 3 and chapter 4 are joint works with Shixin Zheng, Jianfeng Lu, Govind Menon and Xiangxiong Zhang. The two papers concerns the theoretical part and numerical part of the Gibbs sampling problem on the manifold of fixed rank positive definite matrices, where Riemannian Langevin equation is the main tool used to generate desired Gibbs distribution. The main results are two explicit RLE on the manifold as well as two numerical schemes using Euler-Maruyama discretization. The discussion of convergence and verification of equilibrium, as well as the use of the sampling on numerical integral are also included. This work may be interesting to people in the area of optimization and statistics on manifolds.

Chapter 5 is a joint work with Govind Menon. In this work, RLE serves as stochastic relaxation of interior point method in conic programming. Specifically, RLE is discussed on convex cones equipped with Cheng-Yau metric, the Hessian metric generated by canonical barrier. Properties of Brownian motion associated to canonical barrier are proved, and a concrete example of Lorentz cone is discussed in detail. As an extension of classical optimization algorithm, this work may attract people working on optimization. Also, this work provides analysis of Cheng-Yau metric in a probabilistic prospective and may be interesting to people interested in Kahler geometry.

Chapter 6 is a joint work with Govind Menon, where the effect of mean curvature flow of Brownian motion is used to explain the implicit regularization phenomena in deep learning, i.e. even there are multiple optimizer in the parameter space, only a few of them appears in practice without introducing any external force. The model under consideration is Deep Linear Network (DLN) where all active functions are identity, so the system is totally linear and easier to analyze than general deep neutral networks. The product of many matrices is the main object to analyze, which was discussed in random matrix theory literature. This work may be interesting to people working on theory of deep neutral network or random matrix theory.

Chapter 1

The Riemannian Langevin Equation

The Riemannian Langevin Equation on a Riemannian manifold $(\mathcal{M}, g)$ with energy $\mathcal{E} \in C^2(\mathcal{M})$ stands for the following SDE

$$\mathrm{d}\boldsymbol{X}_t = -\mathrm{grad}\,\mathcal{E}(\boldsymbol{X}_t)\mathrm{d}t + \mathrm{d}\boldsymbol{B}_t^{g,\beta} \tag{1.0.1}$$

with $\boldsymbol{B}_t^{g,\beta}$ being Brownian motion on $(\mathcal{M}, g)$ under temperature $\frac{1}{\beta}$. Terms to be defined later, the purpose of this chapter is to go through the preliminaries of the Riemannian Langevin equation, and the main goal is to express Brownian motion as solution to SDEs. Direct constructions can be found in 1.2 and indirect constructions are in 1.3. Based on well-established theories of Brownian motion on Riemannian manifolds, the main contributions in this chapter is 4 where a topological obstacle of expressing Brownian motion using SDE is solved for a class of Riemannian manifolds.

1.1 Background

Main definition about stochastic processes on manifolds follows [43]. Let $(\mathcal{M}, g)$ be a smooth Riemannian manifold, the following definitions are needed: for $f \in C^2(\mathcal{M})$, the Riemannian gradient $\operatorname{grad} f$ is an element in $\mathfrak{X}(\mathcal{M})$, the collection of all vector fields on $\mathcal{M}$, defined by

$$\langle \operatorname{grad} f, \boldsymbol{V} \rangle = \boldsymbol{V} f = \mathrm{d}f(\boldsymbol{V}) \tag{1.1.1}$$

for any $\boldsymbol{V} \in \mathfrak{X}(\mathcal{M})$, where $\langle \cdot, \cdot \rangle$ is the inner product of two vector fields defined by the metric, $\mathrm{d}f(\boldsymbol{V})$ is the contraction of $1-$form $\mathrm{d}f$ and vector field $\boldsymbol{V}$, and $\boldsymbol{V} f$ is the action of $\boldsymbol{V}$ on f, the directional derivative of f along $\boldsymbol{V}$. In local coordinate chart $\{x_i\}_{i=1}^d$, $d = \dim \mathcal{M}$, $\operatorname{grad} f$ is expressed by

$$\operatorname{grad} f = g^{ij} \frac{\partial f}{\partial x^i} \frac{\partial}{\partial x^j} \tag{1.1.2}$$

where $\{\frac{\partial}{\partial x^i}\}_{i=1}^d$ is locally a basis of $\mathfrak{X}(\mathcal{M})$, g^{ij} is the inverse of metric $g_{ij} = \langle \frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^i} \rangle$.

For any (smooth) vector field $\boldsymbol{V} = V^i \frac{\partial}{\partial x^i} \in \mathfrak{X}(\mathcal{M})$, with components $\{V^i\}_{i=1}^d$, the divergence $\operatorname{div}\boldsymbol{V}$ is the function

$$\operatorname{div}\boldsymbol{V} = \frac{1}{\sqrt{g}} \frac{\partial}{\partial x^i} (\sqrt{g}\, V^i) \tag{1.1.3}$$

where $g = \det(g_{ij})$ is the determinant of matrix g_{ij}.

Combining these two operator we get the Laplace-Beltrami operator Δ_g, for $f \in C^2(\mathcal{M})$

$$\Delta_g f = \operatorname{div} \operatorname{grad} f \tag{1.1.4}$$

as the generalization of Laplacian in Euclidean spaces.

Assuming the understanding of Levi-Civita connection ∇ on Riemannian manifolds, the Riemannian Hessian of a function $f \in C^2(\mathcal{M})$ is the symmetric $2-$form field in $\mathfrak{X}(T^*\mathcal{M} \otimes T^*\mathcal{M})$, given by

$$\nabla^2 f = \nabla \mathrm{d} f = (\frac{\partial^2 f}{\partial x^i \partial x^j} - \Gamma^k_{ij} \frac{\partial f}{\partial x^k}) \mathrm{d}x^i \otimes \mathrm{d}x^j \tag{1.1.5}$$

where

$$\Gamma^k_{ij} = \frac{1}{2} g^{kl} (\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij}) \tag{1.1.6}$$

is Christoffel Symbol, and the following relationship holds

$$\Delta_g f = \mathrm{Tr}(\nabla^2 f) = g^{ij} (\frac{\partial^2 f}{\partial x^i \partial x^j} - \Gamma^k_{ij} \frac{\partial f}{\partial x^k}) \tag{1.1.7}$$

With connection ∇ on Riemannian manifold, the concept of martingale can be defined as follows: a $\mathcal{M}$-valued semimartingale $\boldsymbol{X}_t$ is called a ∇-martingale if for any $f \in C^\infty(\mathcal{M})$, the following $\mathbb{R}$-valued process

$$f(\boldsymbol{X}_t) - f(\boldsymbol{X}_0) - \frac{1}{2} \int_0^t \nabla^2 f(\mathrm{d}\boldsymbol{X}_s, \mathrm{d}\boldsymbol{X}_s) \tag{1.1.8}$$

is a local martingale, where $\nabla^2 f$ is the Hessian of f.

Such characterization of martingale on manifolds overcame the problem that there is in general no analogue of linear functions on manifolds which can be used to describe martingales on Euclidean spaces. Also, the definition of martingale depends on connection, meaning that the concept of martingale depends on geometry of the underlying space. Although the connection can be defined independently without the knowledge of the metric, in the following contents only Levi-Civita connection will be considered, as this special connection provides much more information.

This definition of martingale on manifolds is different from the one in [43], but we use this for simplicity. It also worth noting that when $(\mathcal{M}, g)$ is just $\mathbb{R}^n$, this definition of martingale is the parallel of local martingale in Euclidean spaces.

Brownian motion on Riemannian manifolds is similarly defined by the following: a $\mathcal{M}$-valued semimartingale $\boldsymbol{X}_t$ is called a Brownian motion on $(\mathcal{M}, g)$ with temperature $\frac{1}{\beta}$, if for any $f \in C^\infty(\mathcal{M})$, the following $\mathbb{R}$-valued process

$$f(\boldsymbol{X}_t) - f(\boldsymbol{X}_0) - \frac{1}{\beta}\int_0^t \Delta_g f(\boldsymbol{X}_s)\mathrm{d}s \tag{1.1.9}$$

is a local martingale. By the property $\Delta_g = \mathrm{Tr}\nabla^2$, Brownian motion is a $\nabla-$martingale.

On Riemannian manifolds with Levi-Civita connection, the knowledge of Brownian motion is in fact equivalent to the knowledge of the metric, as under local coordinate chart,

$$\Delta_g(x^i x^j) = x^i \Delta_g x^j + x^j \Delta_g x^i + 2\langle \operatorname{grad} x^i, \operatorname{grad} x^j \rangle = x^i \Delta_g x^j + x^j \Delta_g x^i + 2g^{ij} \tag{1.1.10}$$

so the inverse metric g^{ij} is obtained.

1.2 SDE on manifolds

In Euclidean spaces $\mathbb{R}^n$, SDE can be written in both Ito form

$$\mathrm{d}X_t^i = \sum_{j=1}^{n} \sigma_j^i(X_t)\mathrm{d}B_t^j + \mu^i(X_t)\mathrm{d}t \tag{1.2.1}$$

and Stratonovich form

$$\mathrm{d}X_t^i = \sum_{j=1}^{n} \sigma_j^i(X_t) \circ \mathrm{d}B_t^j + b^i(X_t)\mathrm{d}t \tag{1.2.2}$$

with tranformation rule

$$b^i = \mu^i - \sum_{j,k=1}^{n} \frac{1}{2}\sigma_j^k \frac{\partial \sigma_j^i}{\partial x^k} \tag{1.2.3}$$

where $\{\sigma_j^i\}_{i,j=1}^n$, $\{b^i, \mu^i\}_{i=1}^n$ are smooth functions, and $\{B^i\}_{i=1}^n$ are n independent standard Brownian motion.

Being more intuitive, solution to Ito SDE is a local martingale as long as $\mu^i = 0$, with the cost that the evolution of functions follows Ito's formula. The advantage of Stratonovich SDE is that it preserves differential structure, in the sense that for any $f \in C^\infty(\mathbb{R}^n)$,

$$\mathrm{d}f(X_t) = \sum_{i,j=1}^{n} \sigma_j^i \frac{\partial f}{\partial x^i}(X_t) \circ \mathrm{d}B_t^j + \sum_{i=1}^{n} b^i \frac{\partial f}{\partial x^i}(X_t)\mathrm{d}t \tag{1.2.4}$$

On manifolds, in fact, Stratonovich SDE turns out to be more general. Without the assumption of metric structure, on a differential manifold $\mathcal{M}$, a SDE can be written as follows

$$\mathrm{d}\boldsymbol{X}_t = \sum_{\alpha=1}^{l} \boldsymbol{V}_\alpha(\boldsymbol{X}_t) \circ \mathrm{d}Z_t^\alpha \tag{1.2.5}$$

where $\{\boldsymbol{V}_\alpha\}_{\alpha=1}^l \subset \mathfrak{X}(\mathcal{M})$ are l vector fields, $\{Z^\alpha\}_{\alpha=1}^l$ are l $\mathbb{R}$-valued semimartingale. A $\mathcal{M}$-valued process $\boldsymbol{X}_t$ is said to be a solution to this SDE if for any $f \in C^\infty(\mathcal{M})$,

$$f(\boldsymbol{X}_t) = f(\boldsymbol{X}_0) + \int_0^t \sum_{\alpha=1}^{l} \boldsymbol{V}_\alpha f(\boldsymbol{X}_s) \circ \mathrm{d}Z_s^\alpha \tag{1.2.6}$$

It worth noting that in the following content only weak solution is considered, i.e. two solutions $\boldsymbol{X}_t^{(1)}$ and $\boldsymbol{X}_t^{(2)}$ to the SDE are regarded as the same as long as for each $f \in C^\infty(\mathcal{M})$, $f(\boldsymbol{X}_t^{(1)})$ and $f(\boldsymbol{X}_t^{(2)})$ have the same distribution. Existence of solution to SDE will not be discussed here.

The analogue of Ito's SDE on manifolds requires the understanding of connection, and we have the following theorem providing a useful construction of Brownian motion on Riemannian manifolds,

Theorem 1. *On a Riemannian manifold $(\mathcal{M}, g)$ equipped with Levi-Civita connection ∇, given smooth vector fields $\{\boldsymbol{V}_\alpha\}_{\alpha=1}^l$, the solution to the following SDE is a ∇-martingale*

$$\mathrm{d}\boldsymbol{X}_t = \sum_{\alpha=1}^{l} (\boldsymbol{V}_\alpha(\boldsymbol{X}_t) \circ \mathrm{d}B_t^\alpha - \frac{1}{2}\nabla_{\boldsymbol{V}_\alpha}\boldsymbol{V}_\alpha \mathrm{d}t) \tag{1.2.7}$$

where $\{B_t^\alpha\}_{\alpha=1}^l$ are l independent $\mathbb{R}-$valued standard Brownian motions.

Proof. By definition of the solution to SDE, for any $f \in C^\infty(\mathcal{M})$

$$f(\boldsymbol{X}_t) = f(\boldsymbol{X}_t) + \int_0^t \sum_{\alpha=1}^{l} \Big(\boldsymbol{V}_\alpha f(\boldsymbol{X}_s) \circ \mathrm{d}B_s^\alpha - \frac{1}{2}\nabla_{\boldsymbol{V}_\alpha}\boldsymbol{V}_\alpha f(\boldsymbol{X}_s)\mathrm{d}s\Big) \tag{1.2.8}$$

Using the property

$$\boldsymbol{V}(\boldsymbol{V}(f)) = \nabla^2 f(\boldsymbol{V}, \boldsymbol{V}) + \nabla_{\boldsymbol{V}} \boldsymbol{V} f \tag{1.2.9}$$

and transform the Stratonovich integral into Ito integral, we have

$$\begin{aligned} f(\boldsymbol{X}_t) =& f(\boldsymbol{X}_t) + \int_0^t \sum_{\alpha=1}^{l} \Big(\boldsymbol{V}_\alpha f(\boldsymbol{X}_s) \mathrm{d}B_s^\alpha + \frac{1}{2} \boldsymbol{V}_\alpha(\boldsymbol{V}_\alpha(f))(\boldsymbol{X}_s)\mathrm{d}s - \frac{1}{2} \nabla_{\boldsymbol{V}_\alpha} \boldsymbol{V}_\alpha f(\boldsymbol{X}_s)\mathrm{d}s \Big) \\ =& f(\boldsymbol{X}_t) + \int_0^t \sum_{\alpha=1}^{l} \Big(\boldsymbol{V}_\alpha f(\boldsymbol{X}_s) \mathrm{d}B_s^\alpha + \nabla^2 f(\boldsymbol{V}_\alpha, \boldsymbol{V}_\alpha)(\boldsymbol{X}_s)\mathrm{d}s \Big) \end{aligned} \tag{1.2.10}$$

As the integral $\int_0^t \boldsymbol{V}_\alpha f(\boldsymbol{X}_t) \mathrm{d}B_t^\alpha$ is a local martingale, $\boldsymbol{X}_t$ is a ∇-martingale. □

With this theorem, we are able to construct ∇-martingales using SDE on manifolds. In particular, if vector fields $\{\boldsymbol{V}_\alpha\}_{\alpha=1}^l$ forms the inverse metric g^{ij}, we are able to construct Brownian motion on $(\mathcal{M}, g)$.

Theorem 2. *If vector fields $\{\boldsymbol{V}_\alpha\}_{\alpha=1}^l$ satisfies that*

$$\sum_{\alpha=1}^{l} V_\alpha^i V_\alpha^j = g^{ij} \tag{1.2.11}$$

on each local coordinate chart, where $V_\alpha^i = \boldsymbol{V}_\alpha(x^i)$ is the coordinate of $\boldsymbol{V}_\alpha$, then the solution to the following SDE is a Brownian motion on $(\mathcal{M}, g)$ with temperature $\frac{1}{\beta}$

$$\mathrm{d}\boldsymbol{X}_t = \sum_{\alpha=1}^{l} \Big(\sqrt{\frac{2}{\beta}} \boldsymbol{V}_\alpha(\boldsymbol{X}_t) \circ \mathrm{d}B_t^\alpha - \frac{1}{\beta} \nabla_{\boldsymbol{V}_\alpha} \boldsymbol{V}_\alpha(\boldsymbol{X}_t)\mathrm{d}t \Big) \tag{1.2.12}$$

where $\{B_t^\alpha\}_{\alpha=1}^l$ are l independent $\mathbb{R}-$valued standard Brownian motions.

Remark 1. Expressing Brownian motion as a solution to SDE is an old question and it is well known that if the Laplace-Beltrami operator can be expressed as

$$\frac{1}{2}\Delta_g = \frac{1}{2}\sum_{i=1}^{l} \boldsymbol{V}_i^2 + \boldsymbol{V}_0 \tag{1.2.13}$$

then the solution to the SDE

$$\mathrm{d}\boldsymbol{X}_t = \sum_{i=1}^{l} \boldsymbol{V}_i(\boldsymbol{X}_t) \circ \mathrm{d}B_t^i + \boldsymbol{V}_0(\boldsymbol{X}_t)\mathrm{d}t \tag{1.2.14}$$

is a Brownian motion on the manifold $(\mathcal{M}, g)$, where $\{B_t^i\}_{i=1}^l$ are l independent $\mathbb{R}-$valued Brownian motions.

This result is implicitly stated in the work [11]. The relationship between $\boldsymbol{V}_0$ and $\{\boldsymbol{V}_i\}_{i=1}^l$ was also discussed in literature [82] for Brownian motion on Lie groups, and on Riemannian manifolds the results are similar.

Remark 2. The decomposition 1.2.11 is commonly used for construction of Brownian motions and appears in literature including [57, 86]. This theorem is the natural extension of these results to manifolds, while the difference turns out to be some topological restrictions which is partially solved in 4.

Proof. Following the computation above, take any $f \in C^\infty(\mathcal{M})$,

$$\begin{aligned} f(\boldsymbol{X}_t) =& f(\boldsymbol{X}_0) + \int_0^t \sum_{\alpha=1}^{l} \Big(\boldsymbol{V}_\alpha f(\boldsymbol{X}_s)\mathrm{d}B_s^\alpha + \nabla^2 f(\boldsymbol{V}_\alpha, \boldsymbol{V}_\alpha)(\boldsymbol{X}_s)\mathrm{d}s \Big) \\ =& f(\boldsymbol{X}_0) + \int_0^t \sum_{\alpha=1}^{l} \Big(\sqrt{\frac{2}{\beta}} \boldsymbol{V}_\alpha f(\boldsymbol{X}_s)\mathrm{d}B_s^\alpha + \frac{1}{\beta}\Delta_g f(\boldsymbol{X}_s)\mathrm{d}s \Big) \end{aligned} \tag{1.2.15}$$

where we used the property

$$\begin{aligned}\sum_{\alpha=1}^{l}\nabla^2 f(\boldsymbol{V}_\alpha,\boldsymbol{V}_\alpha)=&\sum_{\alpha=1}^{l}(\frac{\partial^2 f}{\partial x^i\partial x^j}-\Gamma_{ij}^k\frac{\partial f}{\partial x^k})V_\alpha^i V_\alpha^j\\ =&(\frac{\partial^2 f}{\partial x^i\partial x^j}-\Gamma_{ij}^k\frac{\partial f}{\partial x^k})g^{ij}\\ =&\Delta_g f\end{aligned}\tag{1.2.16}$$

Therefore, $\boldsymbol{X}_t$ is a Brownian motion with temperature $\frac{1}{\beta}$ on $(\mathcal{M},g)$. □

This theorem provides a way to express Brownian motion on $(\mathcal{M},g)$ as a solution to a SDE. Now the only problem is the existence of such vector fields $\{\boldsymbol{V}_\alpha\}_{\alpha=1}^{l}$. In fact, there are several choices of the vector fields satisfying the condition.

Corollary 3. If $\{\boldsymbol{V}_\alpha\}_{\alpha=1}^{\dim\mathcal{M}}$ is a global orthonormal basis of $\mathfrak{X}(\mathcal{M})$, then the solution to the SDE

$$\mathrm{d}\boldsymbol{X}_t=\sum_{\alpha=1}^{\dim\mathcal{M}}(\sqrt{\frac{2}{\beta}}\boldsymbol{V}_\alpha(\boldsymbol{X}_t)\circ\mathrm{d}B_t^\alpha-\frac{1}{\beta}\nabla_{\boldsymbol{V}_\alpha}\boldsymbol{V}_\alpha\mathrm{d}t)\tag{1.2.17}$$

is a Brownian motion with temperature $\frac{1}{\beta}$ on $(\mathcal{M},g)$, where $\{B_t^\alpha\}_{\alpha=1}^{\dim\mathcal{M}}$ are $\dim\mathcal{M}$ independent $\mathbb{R}-$valued standard Brownian motions.

This corollary requires the existence of a global orthonormal basis of $\mathfrak{X}(\mathcal{M})$, which doesn't hold in general because of some topological restriction. Manifolds having global basis is called parallelizable, and many manifolds are not parallelizable, e.g. $\mathbb{S}^2$. However, using partition of unity we are able to do much more extension.

Corollary 4. If $(\mathcal{M},g)$ has the following finite open cover $\{U_\beta\}_{\beta=1}^{m}$ where each U_β is simply connected, let $\{\rho_\gamma\}_{\gamma=1}^{m}$ be the partition of unity associated to this cover, and $\{\boldsymbol{V}_{\alpha,\gamma}\}_{\alpha=1}^{\dim\mathcal{M}}$ be orthonormal basis locally defined on each U_γ, then for each α,γ, vector field $\sqrt{\rho_\gamma}V_{\alpha,\gamma}$ is globally well-defined, and the solution to the SDE

$$\mathrm{d}\boldsymbol{X}_t = \sum_{\gamma=1}^{m}\sum_{\alpha=1}^{\dim\mathcal{M}}(\sqrt{\frac{2}{\beta}\rho_\gamma}\boldsymbol{V}_{\alpha,\gamma}(\boldsymbol{X}_t)\circ\mathrm{d}B_t^{\alpha,\gamma} - \frac{1}{\beta}\nabla_{\sqrt{\rho_\gamma}\boldsymbol{V}_{\alpha,\gamma}}\sqrt{\rho_\gamma}\boldsymbol{V}_{\alpha,\gamma}\mathrm{d}t) \tag{1.2.18}$$

is a Brownian motion with temperature $\frac{1}{\beta}$ on $(\mathcal{M}, g)$, where $\{B_t^{\alpha,\gamma}|1 \leq \alpha \leq \dim\mathcal{M}, 1 \leq \gamma \leq m\}$ are $m \times \dim\mathcal{M}$ independent $\mathbb{R}$-valued standard Brownian motions.

With this corollary, Brownian motion on a wide class of manifolds can be described as solution to SDE.

In the following contents, we use $\boldsymbol{B}_t^{g,\beta}$ to denote Brownian motion on a Riemannian manifold $(\mathcal{M}, g)$ with temperature $\frac{1}{\beta}$. In a SDE of the form

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{\mu}(\boldsymbol{X}_t)\mathrm{d}t + \mathrm{d}\boldsymbol{B}_t^{g,\beta} \tag{1.2.19}$$

the term $\mathrm{d}\boldsymbol{B}_t^{g,\beta}$ should be understood as the summation

$$\mathrm{d}\boldsymbol{B}_t^{g,\beta} = \sum_{\alpha=1}^{l}\left(\sqrt{\frac{2}{\beta}}\boldsymbol{V}_\alpha(\boldsymbol{X}_t)\circ\mathrm{d}B_t^{\alpha,\beta} - \frac{1}{\beta}\nabla_{\boldsymbol{V}_\alpha}\boldsymbol{V}_\alpha(\boldsymbol{X}_t)\mathrm{d}t\right) \tag{1.2.20}$$

for vector fields satisfying the condition in 2.

1.3 Transition properties of Brownian motion

Brownian motion exhibits geometric properties of the manifolds, the knowledge of Brownian motion is equivalent to the knowledge of metric. In the case $(\mathcal{M}, g)$ is a submanifold or a quotient manifold of another Riemannian manifold $(\tilde{\mathcal{M}}, \tilde{g})$, the images of Brownian motion under corresponding maps are also related to metric obtained by Riemannian immersion and submersion.

In the case $(\mathcal{M}, g)$ is a Riemannian submanifold of $(\tilde{\mathcal{M}}, \tilde{g})$, the metric g is obtained by projecting $\tilde{g}$ onto $T\mathcal{M}$, i.e.

$$g(\boldsymbol{U}, \boldsymbol{V}) = \tilde{g}(\boldsymbol{U}, \boldsymbol{V}) \tag{1.3.1}$$

for $\boldsymbol{U}, \boldsymbol{V} \in (\mathcal{M})$. The Levi-Civita connections ∇ and $\tilde{\nabla}$ are also related by

$$g(\nabla_{\boldsymbol{U}} \boldsymbol{V}, \boldsymbol{W}) = \tilde{g}(\tilde{\nabla}_{\boldsymbol{U}} \boldsymbol{V}, \boldsymbol{W}) \tag{1.3.2}$$

say $\nabla_{\boldsymbol{U}} \boldsymbol{V}$ is also the orthogonal projection of $\tilde{\nabla}_{\boldsymbol{U}} \boldsymbol{V}$ onto $T\mathcal{M}$.

The difference between two connections making a $\nabla-$martingale not necessarily a $\tilde{\nabla}-$martingale. We have the following property

Proposition 5. Let $(\mathcal{M}, g)$ be a Riemannian submanifold of $(\tilde{\mathcal{M}}, \tilde{g})$, with $i : \tilde{\mathcal{M}} \to \mathcal{M}$ the inclusion map. For any $f \in C^\infty(\tilde{\mathcal{M}})$, $\boldsymbol{X} \in \mathcal{M}$, $\boldsymbol{Y} = i(\boldsymbol{X})$, we have

$$\Delta_g(f \circ i)(\boldsymbol{X}) = \mathrm{Tr}|_{T\mathcal{M}}(\tilde{\nabla}^2 f)(\boldsymbol{Y}) + \tilde{\boldsymbol{H}} f(\boldsymbol{Y}) \tag{1.3.3}$$

where $\mathrm{Tr}|_{T\mathcal{M}}$ is the trace in the subspace $T\mathcal{M} \subset T\tilde{\mathcal{M}}$, and $\tilde{\boldsymbol{H}}(\boldsymbol{Y})$ is the mean curvature of $\mathcal{M}$ at $\boldsymbol{Y}$, defined as

$$\tilde{\boldsymbol{H}} = \sum_{i=1}^{\dim\mathcal{M}} (\tilde{\nabla}_{\tilde{\boldsymbol{V}}_i} \tilde{\boldsymbol{V}}_i)^{\perp} \tag{1.3.4}$$

for $\{\tilde{\boldsymbol{V}}_i\}_{i=1}^{\dim\mathcal{M}} \subset \mathfrak{X}(\tilde{\mathcal{M}})$ is an orthonormal basis locally defined in a neighborhood of $\boldsymbol{Y}$, and $(\boldsymbol{U})^{\perp}$ is the projection of $\boldsymbol{U}$ to the normal space $(T\mathcal{M})^{\perp}$.

Proof. This is a point-wise property and we only need to prove this for any $\boldsymbol{X} \in \mathcal{M}$. Take a smooth curve $\gamma(t) : [-1, 1] \to \mathcal{M}$, with $\gamma(0) = \boldsymbol{X}$. Let $\boldsymbol{V}$ be a vector field with $\dot{\gamma}(t) = \boldsymbol{V}(\gamma(t))$. Take derivative of $(f \circ i)(\gamma(t))) = f(i(\gamma(t)))$ at $t = 0$, we have

$$\boldsymbol{V}(f \circ i)(\boldsymbol{X}) = \tilde{\boldsymbol{V}} f(\boldsymbol{Y}) \tag{1.3.5}$$

where $\tilde{\boldsymbol{V}} = i_*(\boldsymbol{V}) \in T\tilde{\mathcal{M}}$ is just $\boldsymbol{V}$ itself, but we use this notation to distinguish with $\boldsymbol{V} \in T\mathcal{M}$. Take secondary derivative we have

$$\nabla_{\boldsymbol{V}} \boldsymbol{V}(f \circ i)(\boldsymbol{X}) + \nabla^2(f \circ i)(\boldsymbol{V}, \boldsymbol{V})(\boldsymbol{X}) = \tilde{\nabla}_{\tilde{\boldsymbol{V}}} \tilde{\boldsymbol{V}} f(\boldsymbol{Y}) + \tilde{\nabla}^2 f(\tilde{\boldsymbol{V}}, \tilde{\boldsymbol{V}})(\boldsymbol{Y}) \tag{1.3.6}$$

Take summation of $\boldsymbol{V}$ over an orthonormal basis $\{\boldsymbol{V}_i\}_{i=1}^{\dim\mathcal{M}}$ locally defined in a neighborhood of $\boldsymbol{X}$, we have

$$\begin{aligned}
\Delta_g(f \circ i)(\boldsymbol{X}) =& \mathrm{Tr}|_{T\mathcal{M}} \tilde{\nabla}^2 f(\boldsymbol{Y}) + \sum_{i=1}^{\dim\mathcal{M}} (\tilde{\nabla}_{\tilde{\boldsymbol{V}}_i} \tilde{\boldsymbol{V}}_i - \nabla_{\boldsymbol{V}_i} \boldsymbol{V}_i) f(\boldsymbol{Y}) \\
=& \mathrm{Tr}|_{T\mathcal{M}} \tilde{\nabla}^2 f(\boldsymbol{Y}) + \sum_{i=1}^{\dim\mathcal{M}} (\tilde{\nabla}_{\tilde{\boldsymbol{V}}_i} \tilde{\boldsymbol{V}}_i)^{\perp} f(\boldsymbol{Y}) \\
=& \mathrm{Tr}|_{T\mathcal{M}} \tilde{\nabla}^2 f(\boldsymbol{Y}) + \tilde{\boldsymbol{H}} f(\boldsymbol{Y})
\end{aligned} \tag{1.3.7}$$

□

The expression of Laplace-Beltrami operator Δ_g above can be used to describe

Brownian motion on $(\mathcal{M}, g)$.

Corollary 6. Let $(\mathcal{M}, g)$ be a submanifold of $(\tilde{\mathcal{M}}, \tilde{g})$ with induced metric, and $i : \mathcal{M} \to \tilde{\mathcal{M}}$ be the inclusion map. Then the image $\boldsymbol{X}_t = i(\boldsymbol{B}_t^{g,\beta})$ of Brownian motion $\boldsymbol{B}_t^{g,\beta}$ is the solution to the following SDE

$$\mathrm{d}\boldsymbol{X}_t = \sum_{\alpha=1}^{l} \Big(\sqrt{\frac{2}{\beta}} \boldsymbol{V}_\alpha(\boldsymbol{X}_t) \circ \mathrm{d}B_t^\alpha - \frac{1}{\beta} \nabla_{\boldsymbol{V}_\alpha} \boldsymbol{V}_\alpha(\boldsymbol{X}_t)\mathrm{d}t \Big) + \frac{1}{\beta} \boldsymbol{H}(\boldsymbol{X}_t)\mathrm{d}t \tag{1.3.8}$$

where $\{B^\alpha\}_{\alpha=1}^l$ are l independent $\mathbb{R}$-valued standard Brownian motion, $\{\boldsymbol{V}_\alpha\}_{\alpha=1}^l \subset \mathfrak{X}(\tilde{\mathcal{M}})$ are vector fields satisfies

$$\sum_{\alpha=1}^{l} V_\alpha^i V_\alpha^j = g^{ij} \tag{1.3.9}$$

and $\boldsymbol{H}(\boldsymbol{X}_t)$ is the mean curvature of $\mathcal{M}$ at $\boldsymbol{X}_t$.

Remark 7. This result is widely known and appears in literature [61] for submanifolds of Euclidean spaces and [85] for submanifolds of general Riemannian manifolds.

Now consider the case where $(\tilde{\mathcal{M}}, \tilde{g})$ has an isometry Lie group G, and $\mathcal{M} = \tilde{\mathcal{M}}/G$ is the quotient manifold, with $\pi : \tilde{\mathcal{M}} \to \mathcal{M}$ the canonical projection map. For each $\boldsymbol{X} \in \mathcal{M}$, the preimage $\pi^{-1}(\boldsymbol{X})$ is a G-orbit in $(\tilde{\mathcal{M}}, \tilde{g})$. The metrics g on $\mathcal{M}$ and $\tilde{g}$ are related through Riemannian submersion as follows. For each $\boldsymbol{X}$ and $\boldsymbol{Y} \in \pi^{-1}(\boldsymbol{X})$, the differential $\pi_* : T_{\boldsymbol{X}}\mathcal{M} \to T_{\boldsymbol{Y}}\tilde{\mathcal{M}}$ is surjective, and the kernel $\mathrm{Ker}(\pi_*)_{\boldsymbol{Y}} \subset T_{\boldsymbol{Y}}\tilde{\mathcal{M}}$ is in general non-empty. Using $\tilde{g}$ we can find the orthogonal complement $\mathrm{Ker}(\pi_*)_{\boldsymbol{Y}}^{\perp} \subset T_{\boldsymbol{Y}}\tilde{\mathcal{M}}$, then restricted to $\mathrm{Ker}(\pi_*)_{\boldsymbol{Y}}^{\perp}$, π_* is a bijection. For each $\boldsymbol{U}, \boldsymbol{V} \in T_{\boldsymbol{X}}\mathcal{M}$, let $\tilde{\boldsymbol{U}}, \tilde{\boldsymbol{V}}$ be their unique preimage in $T_{\boldsymbol{Y}}\tilde{\mathcal{M}}$ respectively, we define

$$g(\boldsymbol{U}, \boldsymbol{V}) = \tilde{g}(\tilde{\boldsymbol{U}}, \tilde{\boldsymbol{V}}) \tag{1.3.10}$$

then π becomes a Riemannian submersion. Notice that as G is an isometry group, this definition of g doesn't depend on the choice of $\boldsymbol{Y} \in \pi^{-1}(\boldsymbol{X})$, so it is well defined.

The Levi-Civita connection $\tilde{\nabla}$ and ∇ are also related as follows, for $\boldsymbol{U}, \boldsymbol{V} \in \mathfrak{X}(\mathcal{M})$,

$$\nabla_{\boldsymbol{U}} \boldsymbol{V} = \pi_*(\tilde{\nabla}_{\tilde{\boldsymbol{U}}} \tilde{\boldsymbol{V}}) \tag{1.3.11}$$

where $\pi_* : T\tilde{\mathcal{M}} \to T\mathcal{M}$ is the differential of $\pi : \tilde{\mathcal{M}} \to \mathcal{M}$, $\tilde{\boldsymbol{U}}, \tilde{\boldsymbol{V}}$ are the unique vector fields in $\mathrm{Ker}(\pi_*)^{\perp}$ such that $\pi_*(\tilde{\boldsymbol{U}}) = \boldsymbol{U}$, $\pi_*(\tilde{\boldsymbol{V}}) = \boldsymbol{V}$.

Similarly, we have the following property

Proposition 8. Let G be an isometry Lie group of $(\tilde{\mathcal{M}}, \tilde{g})$, and $\mathcal{M} = \tilde{\mathcal{M}}/G$ be the quotient manifold, π be the canonical projection. g is defined such that $\pi : (\tilde{\mathcal{M}}, \tilde{g}) \to (\mathcal{M}, g)$ is a Riemannian submersion. The for any $f \in C^{\infty}(\mathcal{M})$, $\boldsymbol{X} \in \mathcal{M}$, $\boldsymbol{Y} \in \pi^{-1}(\boldsymbol{X}) \subset \tilde{\mathcal{M}}$, we have

$$\Delta_{\tilde{g}}(f \circ \pi)(\boldsymbol{Y}) + \tilde{\boldsymbol{H}}(f \circ \pi)(\boldsymbol{Y}) = \Delta_g f(\boldsymbol{X}) \tag{1.3.12}$$

where $\tilde{\boldsymbol{H}}$ is the mean curvature of $\pi^{-1}(\boldsymbol{X})$ regarded as a Riemannian submanifold of $\tilde{\mathcal{M}}$.

In particular, as $\pi^{-1}(\boldsymbol{X})$ is an isometry group orbit, $\boldsymbol{H}(\boldsymbol{X}) = \pi_*(\tilde{\boldsymbol{H}}(\boldsymbol{Y}))$ is a vector field well defined on $\mathcal{M}$, i.e. doesn't depend on the choice of $\boldsymbol{Y} \in \pi^{-1}(\boldsymbol{X})$, the equality can be written as

$$\Delta_{\tilde{g}}(f \circ \pi)(\boldsymbol{Y}) = \Delta_g f(\boldsymbol{X}) - \boldsymbol{H} f(\boldsymbol{X}) \tag{1.3.13}$$

Proof. Take a smooth curve $\tilde{\gamma} : [-1, 1] \to \tilde{\mathcal{M}}$ with $\tilde{\gamma}(0) = \boldsymbol{Y}$, let $\gamma(t) = \pi(\tilde{\gamma}(t))$. $\tilde{\boldsymbol{V}}$ and $\boldsymbol{V}$ are the vector fields such that $\dot{\tilde{\gamma}}(t) = \tilde{\boldsymbol{V}}(\tilde{\gamma}(t))$ and $\dot{\gamma}(t) = \boldsymbol{V}(\gamma(t))$. Take derivative of $f(\gamma(t)) = (f \circ \pi)(\tilde{\gamma}(t))$ at $t = 0$, we have

$$\boldsymbol{V} f(\boldsymbol{X}) = \tilde{\boldsymbol{V}}(f \circ \pi)(\boldsymbol{Y}) \tag{1.3.14}$$

Take secondary derivative, we have

$$\nabla_{\boldsymbol{V}} \boldsymbol{V} f(\boldsymbol{X}) + \nabla^2 f(\boldsymbol{V}, \boldsymbol{V})(\boldsymbol{X}) = \tilde{\nabla}_{\tilde{\boldsymbol{V}}} \tilde{\boldsymbol{V}}(f \circ \pi)(\boldsymbol{Y}) + \tilde{\nabla}^2 (f \circ \pi)(\tilde{\boldsymbol{V}}, \tilde{\boldsymbol{V}})(\boldsymbol{Y}) \tag{1.3.15}$$

Take summation of $\tilde{\boldsymbol{V}}$ over an orthonormal basis $E = \{\tilde{\boldsymbol{V}}_i\}_{i=1}^{\dim \tilde{\mathcal{M}}}$ locally defined in a neighborhood of $\boldsymbol{Y}$, such that $E_1 = \{\tilde{\boldsymbol{V}}_i\}_{i=1}^{\dim \mathcal{M}}$ is an orthonormal basis of $\mathrm{Ker}(\pi_*)^{\perp}$, then $E_2 = \{\tilde{\boldsymbol{V}}_i\}_{i=\dim \mathcal{M}+1}^{\dim \tilde{\mathcal{M}}}$ is an orthonormal basis of $\mathrm{Ker}(\pi_*)$. For $\tilde{\boldsymbol{V}} \in E_2$, $\pi_*(\tilde{\boldsymbol{V}}) = 0$, so the L.H.S is zero. For $\tilde{\boldsymbol{V}} \in E_1$, we have $\nabla_{\boldsymbol{V}} \boldsymbol{V} = \pi_*(\tilde{\nabla}_{\tilde{\boldsymbol{V}}} \tilde{\boldsymbol{V}})$, therefore

$$\Delta_g f(\boldsymbol{X}) = \sum_{\tilde{\boldsymbol{V}} \in E_2} \tilde{\nabla}_{\tilde{\boldsymbol{V}}} \tilde{\boldsymbol{V}}(f \circ \pi)(\boldsymbol{Y}) + \Delta_{\tilde{g}}(f \circ \pi)(\boldsymbol{Y}) \tag{1.3.16}$$

Since $f \circ \pi$ is a constant over $\pi^{-1}(\boldsymbol{X})$, only the part of $\tilde{\nabla}_{\tilde{\boldsymbol{V}}} \tilde{\boldsymbol{V}}$ in $\mathrm{Ker}(\pi_*)^{\perp}$ =

$T(\pi^{-1}(\boldsymbol{X}))^{\perp}$ makes non-zero contribution, thus

$$
\begin{aligned}
&\sum_{\tilde{\boldsymbol{V}}\in E_2} \tilde{\nabla}_{\tilde{\boldsymbol{V}}}\tilde{\boldsymbol{V}}(f\circ\pi)(\boldsymbol{Y})\\
=&\sum_{\tilde{\boldsymbol{V}}\in E_2} \left(\tilde{\nabla}_{\tilde{\boldsymbol{V}}}\tilde{\boldsymbol{V}}\right)^{\perp}(f\circ\pi)(\boldsymbol{Y})\\
=&\tilde{\boldsymbol{H}}(f\circ\pi)(\boldsymbol{Y})
\end{aligned}
\tag{1.3.17}
$$

□

This proposition also leads to the description of Brownian motion

Corollary 9. Let $(\tilde{\mathcal{M}},\tilde{g})$ and $(\mathcal{M},g)$ be Riemannian manifolds and $\pi:\tilde{\mathcal{M}}\to\mathcal{M}$ is a Riemannian submersion. Let $\boldsymbol{B}_t^{\tilde{g},\beta}$ be the Brownian motion on $\tilde{\mathcal{M}}$ and $\boldsymbol{X}_t=\pi(\boldsymbol{B}_t^{\tilde{g},\beta})$ be its image in $\mathcal{M}$, then $\boldsymbol{X}_t$ is a solution to the following SDE

$$
\mathrm{d}\boldsymbol{X}_t=\sum_{\alpha=1}^{l}(\sqrt{\frac{2}{\beta}}\boldsymbol{V}_\alpha(\boldsymbol{X}_t)\circ\mathrm{d}B_t^\alpha-\frac{1}{\beta}\nabla_{\boldsymbol{V}_\alpha}\boldsymbol{V}_\alpha(\boldsymbol{X}_t)\mathrm{d}t)-\frac{1}{\beta}\boldsymbol{H}(\boldsymbol{X}_t)\mathrm{d}t \tag{1.3.18}
$$

where $\{B_t^\alpha\}_{\alpha=1}^l$ are l independent $\mathbb{R}-$valued standard Brownian motions, $\{\boldsymbol{V}_\alpha\}_{\alpha=1}^l$ satisfies

$$
\sum_{\alpha=1}^{l}V_\alpha^iV_\alpha^j=g^{ij} \tag{1.3.19}
$$

and $\boldsymbol{H}(\boldsymbol{X})=\pi_*(\tilde{\boldsymbol{H}}(\boldsymbol{Y}))$ for $\boldsymbol{Y}\in\pi^{-1}(\boldsymbol{X})$, where $\tilde{\boldsymbol{H}}$ is the mean curvature of $\pi^{-1}(\boldsymbol{X})\subset\tilde{\mathcal{M}}$ at $\boldsymbol{Y}$.

Remark 10. This result is first found in [49] and seems to be rarely known. The condition that the mean curvature vector is a constant on the fiber seems to be deliberate, but is natural in the case that the Riemannian submersion comes from taking quotient of a Riemannian manifold with respect to an isometry group.

For the vector field $\boldsymbol{H}$, the following proposition connects it with G-group orbit volume [77]

Proposition 11. Let $(\tilde{\mathcal{M}}, \tilde{g})$ with isometry group G. Let $\mathcal{M} = \tilde{\mathcal{M}}/G$ and $\pi : \tilde{\mathcal{M}} \to \mathcal{M}$ be canonical projection. g is defined such that π is a Riemannian submersion. Define $\Omega(\boldsymbol{X}) = \mathrm{vol}(\pi^{-1}(\boldsymbol{X}))$ to be the volume of $\pi^{-1}(\boldsymbol{X})$, then the mean curvature $\tilde{\boldsymbol{H}}$ of group orbit $\pi^{-1}(\boldsymbol{X})$ at $\boldsymbol{Y} \in \pi^{-1}(\boldsymbol{X})$, and the projection vector field $\boldsymbol{H}$ satisfies

$$\tilde{\boldsymbol{H}}(\boldsymbol{Y}) = -\,\tilde{\mathrm{grad}}\,(S \circ \pi)(\boldsymbol{Y}) \tag{1.3.20}$$

$$\boldsymbol{H}(\boldsymbol{X}) = -\,\mathrm{grad}\, S(\boldsymbol{X}) \tag{1.3.21}$$

where $S(\boldsymbol{X}) = \log \Omega(\boldsymbol{X})$, grad and $\tilde{\mathrm{grad}}$ are the gradient operator in $(\mathcal{M}, g)$ and $(\tilde{\mathcal{M}}, \tilde{g})$ respectively.

Combine these two, we have the following

Proposition 12. Let $(\tilde{\mathcal{M}}, \tilde{g})$ and $(\mathcal{M}, g)$ be Riemannian manifolds and $i : \tilde{\mathcal{M}} \to \mathcal{M}$ is a Riemannian submersion. Let $\boldsymbol{B}_t^{\tilde{g},\beta}$ be the Brownian motion on $\tilde{\mathcal{M}}$ and $\boldsymbol{X}_t = \pi(\boldsymbol{B}_t^{\tilde{g},\beta})$ be its image in $\mathcal{M}$, then $\boldsymbol{X}_t$ is a solution to the following SDE

$$\mathrm{d}\boldsymbol{X}_t = \frac{1}{\beta}\mathrm{grad}\, S(\boldsymbol{X}_t)\mathrm{d}t + \mathrm{d}\boldsymbol{B}_t^{g,\beta} \tag{1.3.22}$$

where $S(\boldsymbol{X}) = \log \mathrm{vol}(\pi^{-1}(\boldsymbol{X}))$

Notice that the function $S(\boldsymbol{X}) = \log \mathrm{vol}(\pi^{-1}(\boldsymbol{X}))$ coincides with the definition of Boltzmann entropy $S = \log \Omega$, we also call $S(\boldsymbol{X})$ an entropy function. The SDE (12) could be explained as, the image of Brownian motion is another Brownian motion drifted by entropy gradient.

1.4 Riemannian Langevin Equation

Inspired by the SDE (12), we define Riemannian Langevin Equation (RLE) as follows

$$\mathrm{d}\boldsymbol{X}_t = -\mathrm{grad}\,\mathcal{E}(\boldsymbol{X}_t)\mathrm{d}t + \mathrm{d}\boldsymbol{B}_t^{g,\beta} \tag{1.4.1}$$

where $\mathcal{E} \in C^2(\mathcal{M})$ is an energy function.

As extension of Langevin equation in Euclidean spaces, the following propositions are all parallel

Proposition 13. [Fokker-Planck Equation] The density ρ_t of $\boldsymbol{X}_t$ satisfies the following PDE

$$\partial_t \rho_t = \mathrm{div}\Big(\frac{1}{\beta}\mathrm{grad}\,\rho_t + \rho_t \mathrm{grad}\,\mathcal{E}\Big) \tag{1.4.2}$$

Proof. For any $f \in C^\infty(\mathcal{M})$, the evolution of $f(\boldsymbol{X}_t)$ is

$$\mathrm{d}f(\boldsymbol{X}_t) = (\frac{1}{\beta}\Delta_g f(\boldsymbol{X}_t) - \mathrm{grad}\,\mathcal{E} f(\boldsymbol{X}_t)\Big)\mathrm{d}t + \text{local martingale} \tag{1.4.3}$$

so the generator of $\boldsymbol{X}_t$ is $\mathcal{L} = \frac{1}{\beta}\Delta_g - \mathrm{grad}\,\mathcal{E}$, then the dual

$$\mathcal{L}^* \rho = \frac{1}{\beta}\Delta_g \rho + \mathrm{div}(\rho\,\mathrm{grad}\,\mathcal{E}) \tag{1.4.4}$$

and ρ_t satisfies

$$\partial_t \rho_t = \mathcal{L}^* \rho_t = \frac{1}{\beta}\Delta_g \rho_t + \mathrm{div}(\rho_t\,\mathrm{grad}\,\mathcal{E}) \tag{1.4.5}$$

□

It is easy to see that $\rho_* \propto \mathrm{e}^{-\beta\mathcal{E}}$ is the equilibrium measure of (13). To make sure ρ_* is well-defined, we require that

$$Z_\beta := \int_{\mathcal{M}} \mathrm{e}^{-\beta\mathcal{E}} \mathrm{d}V < \infty \tag{1.4.6}$$

where $\mathrm{d}V = \sqrt{g}\mathrm{d}x_1 \cdots \mathrm{d}x_n$, $g = \det(g_{ij})$ is the volume form, so $\rho_* = \frac{1}{Z_\beta}\mathrm{e}^{-\beta\mathcal{E}}$ defines a probability measure on $\mathcal{M}$.

In this case, we have the following convergence property

Proposition 14. Suppose $\mathcal{E}$ satisfies that $Z_\beta < \infty$, take any function $\phi \in C^2(\mathbb{R}_+)$ such that $\phi(\cdot) \geq \phi(1) = 0, \phi''(\cdot) > 0$, then $\mathcal{F}(\rho) := \int_{\mathcal{M}} \phi(\frac{\rho}{\rho_*})\rho_* \mathrm{d}V$ defines a Lyapunov functional for (13). If the initialization satisfies $F(\rho_0) < \infty$, we have the convergence $\lim_{t\to\infty} \mathcal{F}(\rho_t) = 0$.

Proof. Using (13) we get

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(\rho_t) =& \frac{1}{\beta}\int_{\mathcal{M}} \phi'(\frac{\rho_t}{\rho_*})\mathrm{div}(\rho_* \mathrm{grad}\,\frac{\rho_t}{\rho_*})\mathrm{d}V \\
=& -\frac{1}{\beta}\int_{\mathcal{M}} \rho_*\phi''(\frac{\rho}{\rho_*})|\mathrm{grad}\,\frac{\rho}{\rho_*}|_g^2 \mathrm{d}V \leq 0
\end{aligned} \tag{1.4.7}$$

where $|\mathrm{grad}\, f|_g^2 = g^{ij}\partial_i f \partial_j f$ is the norm of $\mathrm{grad}\, f$. The equality holds only if $\mathrm{grad}\,\frac{\rho}{\rho_*} = 0$, i.e. $\rho = \rho_*$, so we are done. ☐

Corollary 15. When $\mathcal{E}$ satisfies $Z_\beta < \infty$, we have the convergence of $\frac{\rho_t}{\rho_*} \to 1$ in $L^2(\rho_* \mathrm{d}V)$, $\rho_t \to \rho_*$ in KL-divergence, i.e. take $\phi(x) = (x-1)^2$ and $\phi(x) = x\log x - x + 1$, we have

$$\lim_{t\to\infty} \int_{\mathcal{M}} \rho_*(\frac{\rho_t}{\rho_*} - 1)^2 \mathrm{d}V = 0 \tag{1.4.8}$$

$$\lim_{t\to\infty} \int_{\mathcal{M}} \rho_t \log\frac{\rho_t}{\rho_*} \mathrm{d}V = 0 \tag{1.4.9}$$

Consider the case where $(\tilde{\mathcal{M}}, \tilde{g})$ has an isometry group G and $\tilde{\mathcal{E}}$ is also invariant under G, i.e.

$$\tilde{\mathcal{E}}(h \circ \boldsymbol{Y}) = \tilde{\mathcal{E}}(\boldsymbol{Y}) \tag{1.4.10}$$

for $\boldsymbol{Y} \in \tilde{\mathcal{M}}, h \in G$, and $h \circ \boldsymbol{Y}$ is the group action. Let the quotient manifold $\mathcal{M} = \tilde{\mathcal{M}}/G$ equipped with metric g such that the canonical projection π is a Riemannian submersion.

Proposition 16. In the settings above, if $\boldsymbol{Y}_t$ satisfies the RLE

$$\mathrm{d}\boldsymbol{Y}_t = -\operatorname{grad} \tilde{\mathcal{E}} + \mathrm{d}\boldsymbol{B}_t^{\tilde{g},\beta} \tag{1.4.11}$$

then $\boldsymbol{X}_t = \pi(\boldsymbol{Y}_t)$ is a solution the the RLE

$$\mathrm{d}\boldsymbol{X}_t = -\operatorname{grad} \mathcal{F} + \mathrm{d}\boldsymbol{B}_t^{g,\beta} \tag{1.4.12}$$

where $\mathcal{F} = \mathcal{E} - \frac{1}{\beta} S$, $\mathcal{E} \in C^2(\mathcal{M})$ satisfies $\tilde{\mathcal{E}} = \mathcal{E} \circ \pi$, and $S(\boldsymbol{X}) = \log \operatorname{vol}(\pi^{-1}(\boldsymbol{X}))$.

The shift of energy function in Riemannian Langevin equation makes the main difference of stochastic relaxation from gradient descent. Variables that have no effect on target function and geometry still affect the process of optimization.

Chapter 2

Siegel Brownian motion

2.0 Abstract

We construct an analogue of Dyson Brownian motion in the Siegel half-space $\mathcal{H}$ that we term *Siegel Brownian motion*. Given $\beta \in (0, \infty]$, a stochastic flow for $Z_t \in \mathcal{H}$ is introduced so that the law of the eigenvalues λ_t of the cross ratio matrix $\mathfrak{R}(Z_t, \boldsymbol{i}I_n)$ is determined, after a change of variables to $\sigma(\lambda) \in (0, \infty)^n$, by the Itô differential equation

$$\mathrm{d}\sigma_t^k = \frac{1}{2}\left(\coth \sigma_t^k + \sum_{l \neq k} \frac{\sinh \sigma_t^k}{\cosh \sigma_t^k - \cosh \sigma_t^l}\right)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}W_t^k, \quad k = 1, \ldots, n, \quad (2.0.1)$$

where W_t is a standard Wiener process in $\mathbb{R}^n$.

This interacting particle system corresponds to stochastic gradient ascent

$$\mathrm{d}\sigma_t = \frac{1}{2}\nabla S(\sigma_t) + \sqrt{\frac{2}{\beta}}\mathrm{d}W_t, \quad (2.0.2)$$

where $S(\sigma) = \log \operatorname{vol} \mathcal{O}_{\lambda(\sigma)}$ is a Boltzmann entropy that enumerates the microstates in the group orbit $\mathcal{O}_\lambda = \{Z \in \mathcal{H} \,|\, \mathrm{eig}\,(\mathfrak{R}(Z, \boldsymbol{i}I_n)) = \lambda\}$. In the limit $\beta = \infty$, the group

orbits $\mathcal{O}_{\lambda_t}$ evolve by motion by minus a half times mean curvature.

2.1 Introduction

2.1.1 Outline

In a recent paper, the first author along with Huang and Inauen, introduced a geometric construction of Dyson Brownian motion [44]. This paper uses this framework to derive an analogue of Dyson Brownian motion in a symmetric space with negative curvature. We focus on the Siegel half-space $\mathcal{H}$ for clarity. While the method can be generalized, it is best understood through this example.

The space $\mathcal{H}$ consists of $n \times n$ complex symmetric matrices whose imaginary part is positive definite, written

$$\mathcal{H} = \left\{ Z = X + \boldsymbol{i}\, Y \middle| X = X^T,\ Y = Y^T,\ Y \succ 0 \right\}. \tag{2.1.1}$$

It is equipped with a metric $g_{\mathcal{H}}$ expressed in terms of the 1-forms dZ as follows

$$g_{\mathcal{H}} = \mathrm{Tr}(Y^{-1}\mathrm{d}Z\, Y^{-1}\mathrm{d}\bar{Z}). \tag{2.1.2}$$

This space was first studied in depth by Siegel [83]. He showed that $\mathcal{H}$ is a symmetric space with negative curvature that generalizes the Poincaré upper half-plane [83, Thms 1-4]. Möbius transformations of the Poincaré upper half-plane, $w = (az+b)/(cz+d)$, where $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{R})$, are replaced by automorphisms of $\mathcal{H}$ of the form

$$W = (AZ+B)(CZ+D)^{-1}. \tag{2.1.3}$$

Here the $n \times n$ real matrices A, B, C, D correspond to elements of the symplectic group $\mathrm{Sp}(n,\mathbb{R})$ as follows. They are the elements of block matrices M such that

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \quad M^T J M = J, \quad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}, \tag{2.1.4}$$

where I is the $n \times n$ identity matrix.

Siegel characterized automorphisms of $\mathcal{H}$ that map a given pair of points (Z, Z_1) into a pair (W, W_1) by introducing the matrix-valued cross-ratio

$$\mathfrak{R}(Z, Z_1) = (Z - Z_1)(Z - \bar{Z}_1)^{-1}(\bar{Z} - \bar{Z}_1)(\bar{Z} - Z_1)^{-1}. \tag{2.1.5}$$

The invariants of the group action (2.1.3) are the eigenvalues of $\mathfrak{R}(Z, Z_1)$: there is an automorphism of $\mathcal{H}$ mapping (Z, Z_1) into (W, W_1) if and only if $\mathfrak{R}(Z, Z_1)$ and $\mathfrak{R}(W, W_1)$ have the same eigenvalues [83, Thm. 2]. These eigenvalues always lie within the interval $[0, 1]$ (a proof is presented in Section 2.3).

We construct a stochastic process $Z_t \in \mathcal{H}$ such that if $\lambda_t = (\lambda_t^1, \dots, \lambda_t^n)$ are the eigenvalues of $\mathfrak{R}(Z_t, iI)$, and $\sigma_t^k > 0$ is the positive solution to the implicit equation

$$\tanh^2 \frac{\sigma_t^k}{2} = \lambda_t^k, \quad 1 \le k \le n, \tag{2.1.6}$$

then $\sigma_t = (\sigma_t^1, \dots, \sigma_t^n)$ has the same law as the unique strong solution to the Itô differential equation

$$\mathrm{d}\sigma_t^k = \frac{1}{2}(\coth \sigma_t^k + \sum_{l \neq k} \frac{\sinh \sigma_t^k}{\cosh \sigma_t^k - \cosh \sigma_t^l})\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}W_t^k, \quad k = 1, \dots, n. \tag{2.1.7}$$

Here $W_t = (W_t^1, \dots, W_t^n)$ is a standard Wiener process in $\mathbb{R}^n$.

Equation (2.1.7) is an analogue of Dyson Brownian motion because they both correspond to stochastic gradient ascent in a geometric sense. For these reasons, we call the interacting particle system in equation (2.1.7) *Siegel Brownian motion*.

Let us now introduce geometric stochastic gradient ascent through examples. We then state two theorems that describe this structure for Siegel Brownian motion. The underlying well-posedness theory parallels that of Dyson Brownian motion. We show in Section 2.5 that equation (2.1.7) has global strong solutions for $\beta \geq 2$.

2.1.2 Motion by curvature and stochastic flows

Let $z \in \mathbb{R}^n$, let $r = |z|$ and let P_z and $P_z^\perp$ denote the orthonormal projections onto the tangent and normal space to the circle S_r^{n-1} of radius r at the point z. Let B_t be a standard Wiener processes in $\mathbb{R}^n$ and consider the Itô SDE

$$\mathrm{d}z_t = P_{z_t}\,\mathrm{d}B_t + \sqrt{\frac{2}{\beta}} P_{z_t}^\perp\,\mathrm{d}B_t. \tag{2.1.8}$$

This SDE is designed so that the point z_t evolves because of independent tangential and normal stochastic fluctuations with the parameter $\beta > 0$ describing the anisotropy of these fluctuations. In the limit $\beta = \infty$, there are only tangential fluctuations and we use Itô's formula to observe that

$$\mathrm{d}(|z_t|^2) = \mathrm{d}(z_t^T z_t) = \mathrm{d}z_t^T z_t + z_t^T \mathrm{d}z_t + \mathrm{d}z_t^T \mathrm{d}z_t = (n-1)\,\mathrm{d}t. \tag{2.1.9}$$

It follows that the radius r_t evolves deterministically by

$$\dot{r}_t = \frac{n-1}{2r_t}. \tag{2.1.10}$$

Thus, when $\beta = \infty$ each point z_t evolves tangentially and stochastically, but the spheres $S_{r_t}^{n-1}$ evolve deterministically by motion by (minus a half times) curvature.

For arbitrary $\beta \in (0, \infty]$, a similar calculation yields that the radius r_t has the same law as the solution to the SDE

$$\mathrm{d}r_t = \frac{n-1}{2r_t}\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}W_t = \frac{1}{2}\nabla S(r_t)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}W_t. \tag{2.1.11}$$

Here W_t is a standard one-dimensional Wiener process, and

$$S(r) = \log \mathrm{vol}\,(S_r^{n-1}),$$

with $\mathrm{vol}(S^{n-1}(r))$ denoting the $n-1$ dimensional volume of the sphere S_r^{n-1}, is the Boltzmann entropy. Observe again that the diffusion of r_t is due to the normal fluctuations driven by B_t, whereas the tangential fluctuations in z_t give rise to the *normal* drift through the Itô correction.

2.1.3 General principles

This example can be systematized through the use of Riemannian submersion. Euclidean space $\mathbb{R}^n$ is foliated by concentric circles of radius $r > 0$ centered at the origin. The space of radii, $r \in [0, \infty)$, may be identified with the quotient space $\mathbb{R}^n/\mathrm{O}(n)$ and equipped with a metric via Riemannian submersion. When constructing the SDE (2.1.8), we introduced stochastic forcing upstairs, with independent horizontal and vertical fluctuations and an anisotropy parameter $\beta > 0$.

In a recent paper [44], this approach was used to construct Dyson Brownian motion at inverse temperature $\beta \in (0, \infty]$. In this case, the space upstairs is the space of Hermitian matrices equipped with the Frobenius norm, writen $(\mathbb{H}(n), \mathrm{Fr})$; the space downstairs is the Weyl chamber of eigenvalues; and the foliation corresponds to isospectral matrices generated by the unitary group [44, §2.1]. Each matrix $M \in \mathbb{H}(n)$ may be diagonalized, written $M = U\Lambda U^*$, so that it lies on a $U(n)$ orbit $\mathcal{O}_\Lambda \subset \mathrm{Her}(n)$. We again use the projections P_M and $P_M^\perp$ onto the tangent and normal spaces to $T_M\mathcal{O}_\Lambda$

to define the Itô differential equation

$$dM_t = P_{M_t} dX_t + \sqrt{\frac{2}{\beta}} P_{M_t}^{\perp} dX_t, \tag{2.1.12}$$

where X_t is the standard Wiener process in $(\mathbb{H}(n), \mathrm{Fr})$.

Equation (2.1.12) is a model for Dyson Brownian motion in the following sense. It is shown in [44, Thm. 1] that the eigenvalues of M_t, written $\lambda_t = (\lambda_t^1, \ldots, \lambda_t^n)$, have the same law as the unique strong solution to the Itô differential equation

$$d\lambda_t^k = \sum_{j \neq k} \frac{1}{\lambda_k - \lambda_j} \, dt + \sqrt{\frac{2}{\beta}} dW_t^k, \quad 1 \leq k \leq n. \tag{2.1.13}$$

Here W_t denotes the standard Wiener process in $\mathbb{R}^n$ and we assume, given $M_0 \in \mathbb{H}(n)$, that $\lambda_0 \in \mathbb{R}^n$ is the spectrum of M_0 presented in increasing order.

Equation (2.1.13) may also be written as

$$d\lambda_t = \frac{1}{2} \nabla S(\lambda) \, dt + \sqrt{\frac{2}{\beta}} dW_t, \quad S(\lambda) = \log \mathrm{vol}(\mathcal{O}_\lambda). \tag{2.1.14}$$

The gradient in λ is with respect to the usual Euclidean metric. This is the metric on $\mathbb{R}^n$ obtained by pushing forward the Frobenius metric on $\mathbb{H}(n)$ under the map that sends a matrix to its eigenvalues. This is the natural Riemannian submersion in this problem. Again, in the limit $\beta = \infty$, the group orbits $\mathcal{O}_{\lambda_t}$ evolve deterministically by motion by minus a half times mean curvature.

2.1.4 A new matrix model

Dyson Brownian motion is the fundamental interacting particle system in random matrix theory. Despite its extensive study, the geometric properties revealed by the construction in [44] have not been noticed before. The simplicity of equation (2.1.12) suggests that other interesting particle systems may be generated through this method. This viewpoint motivated the construction in this paper.

The use of the Siegel half-space allows us to illustrate our method in a symmetric space with negative curvature. In order to do so, we must replace each term in equation (2.1.12) by a geometric analogue.

First, the driving process X_t in equation (2.1.12) is replaced by Browian motion, denoted $\boldsymbol{B}_t$, in $(\mathcal{H}, g)$. This process may be constructed in several ways, as explained in Section 2.2 below; we follow [43, 47].

The analogue of the foliation of $\mathbb{R}^n$ by concentric circles must be provided by a foliation of $\mathcal{H}$ using a suitable group of isometries (observe that an $O(n)$ action generates the concentric circles). We fix the point iI_n as the origin in $\mathcal{H}$ so that its stabilizer subgroup $G_{\boldsymbol{i}I_n} < \mathrm{Sp}(n, \mathbb{R})$ provides the desired group of isometries. Let $\pi : \mathcal{H} \to \mathcal{H}/G_{\boldsymbol{i}I_n}$ be the canonical projection from $\mathcal{H}$ to the quotient space. By Siegel's theorem, the moduli space under this group action may be parametrized by the eigenvalues of the cross-ratio $\mathfrak{R}(Z, \boldsymbol{i}I_n)$. Precisely, the quotient space is

$$\mathcal{H}/G_{\boldsymbol{i}I_n} = \{\lambda \in [0,1)^n | \lambda_1 \leq \lambda_2 \ldots \leq \lambda_n\} . \tag{2.1.15}$$

To make the quotient space a manifold it is necessary to avoid singularities at the boundary, and we must ensure the motion remains in

$$\mathcal{Q} = \mathrm{Int}(\mathcal{H}/G_{\boldsymbol{i}I_n}) = \{\lambda \in (0,1)^n | \lambda_1 < \lambda_2 \ldots < \lambda_n\} . \tag{2.1.16}$$

This space is also equivalent to the quotient of a subset $\tilde{\mathcal{H}} \subset \mathcal{H}$ by $G_{\boldsymbol{i}I_n}$

$$\tilde{\mathcal{H}} = \{\, Z \in \mathcal{H} \,|\, \mathfrak{R}(Z, \boldsymbol{i}I_n) \text{ has no zero or repeated eigenvalues } \}, \quad \mathcal{Q} = \tilde{\mathcal{H}}/G_{\boldsymbol{i}I_n} .$$

It is clear that $\mathcal{Q}$ is a manifold. We equip it with a metric using Riemannian submersion as described below.

For each $\lambda \in \mathcal{Q}$, define the subset of $\mathcal{H}$

$$\mathcal{O}_\lambda = \{Z \in \mathcal{H} \mid \operatorname{eig} \mathfrak{R}(Z, \boldsymbol{i} I_n) = \lambda\}. \tag{2.1.17}$$

Then $\mathcal{O}_\lambda$ is a smooth manifold immersed in $\mathcal{H}$ and it is the $G_{\boldsymbol{i}I_n}$ group orbit of any Z in it. For each $Z \in \mathcal{O}_\lambda$ we then define the orthogonal projection operators $P_Z, P_Z^\perp$ to the tangent space $T_Z\mathcal{O}_\lambda$ and its orthogonal complement $T_Z\mathcal{O}_\lambda^\perp$ respectively.

Fix $\beta \in (0, \infty]$ and assume given $\boldsymbol{Z}_0$ such that the eigenvalues of $\mathfrak{R}(\boldsymbol{Z}_0, i\, I)$ are distinct and lie in $(0, 1)$. We consider the (formal) Itô stochastic differential equation

$$\mathrm{d}\boldsymbol{Z}_t = P_{Z_t}(\mathrm{d}\boldsymbol{B}_t) + \sqrt{\frac{2}{\beta}} P_{Z_t}^\perp(\mathrm{d}\boldsymbol{B}_t), \quad \boldsymbol{Z}(0) = \boldsymbol{Z}_0. \tag{2.1.18}$$

This expression is only formal because stochastic differential equations in $\mathcal{H}$ must be defined using the Stratonovich rule to ensure coordinate invariance. The precise definition of equation (2.1.18), as well as its well-posedness theory, is presented in Section 2.2.2 and Section 2.5. We define the stopping time τ_β as the first time at which eigenvalues to $\mathfrak{R}(Z_t, \boldsymbol{i}I)$ collide. In particular, $\tau_\beta = +\infty$ for $\beta \geq 2$. We then have the following

Theorem 3. *Let $\boldsymbol{Z}_t$ be the unique strong solution to the SDE* (2.1.18)*, let $\lambda_t = \operatorname{eig}(\mathfrak{R}(\boldsymbol{Z}_t, \boldsymbol{i}I_n))$ and define σ_t using equation* (2.1.6)*. Then σ_t has the same law as the unique strong solution to the SDE* (2.1.7) *on the maximal interval of existence $[0, \tau_\beta)$.*

Define the Boltzmann entropy of the group orbit $\mathcal{O}_\lambda$ by the formula

$$S(\lambda) = \log \mathrm{vol}(\mathcal{O}_\lambda), \tag{2.1.19}$$

where the volume form is obtained from the metric g on $\mathcal{H}$. We changed variables from λ to σ in equation (2.1.6) in order that the metric in σ-space is the standard Euclidean metric on $\mathbb{R}^n$. Thus, $\nabla S(\sigma)$ in equation (2.1.21) below denotes the usual Euclidean gradient of the entropy $S(\sigma)$.

Theorem 4. *There is a universal constant $c_n > 0$ such that*

$$S(\sigma) = \sum_{k=1}^{n} \log \sinh \sigma^k + \sum_{1\leq k<l\leq n} \log |\cosh \sigma^k - \cosh \sigma^l| + c_n. \tag{2.1.20}$$

Equation (2.1.7) *is equivalent to stochastic gradient ascent of Boltzmann entropy*

$$\mathrm{d}\sigma_t = \frac{1}{2}\nabla S(\sigma_t)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}W_t. \tag{2.1.21}$$

Remark 5. We have adopted the Boltzmann sign convention for entropy: it is chosen so that our systems evolve to maximize entropy. Conceptually this corresponds to maximizing disorder in the gauge group in each of our examples.

Remark 6. We use the convention in geometric analysis to define the mean curvature of an immersion $u : \mathcal{N} \to (\mathcal{M}, g)$ of a smooth manifold $\mathcal{N}$ into a Riemannian manifold $(\mathcal{M}, g)$): it is the trace of the second fundamental form of u with respect to the metric g. As a guide, it is helpful to note that the mean curvature of a sphere S_r^{n-1} points inwards and it has magnitude $(n-1)/r$.

Remark 7. The mean curvature is usually obtained as the first variation of a volume functional. However, for Riemannian submersion of group orbits generated by isometries, the mean curvature is also the negative of the gradient of the entropy (evaluated upstairs) [78, p.3350]. This is why in all the examples above the group orbits move by motion by minus a half curvature in the limit $\beta = \infty$.

Remark 8. It is well known that Dyson Brownian motion may be generalized using representation theory to the Dunkl process and radial Heckman-Opdam process (see [45, §2] for a comprehensive treatment of these processes, including a history and explicit examples). It is natural to expect a relationship between these processes and Siegel Brownian motion (compare for example equation (2.1.7) and [45, Eq. 2.35]). However, we have been unable to determine such a link. Our approach makes no use of representation theory beyond what is implicit in Siegel's classical work on $\mathcal{H}$; instead our tools are a combination of standard stochastic differential geometry and facts from [83].

The following section provides background material about SDE on manifolds. This is followed by the proofs of Theorem 3 and Theorem 4. The well-posedness theory is described in Section 2.5.

2.2 Background: SDE and martingales on manifolds

2.2.1 Outline

In this section, we assume given a smooth manifold $\mathcal{M}$ (a differentiable manifold with a C^∞ structure) and we review some facts about SDE and martingales on manifolds. The main objective is to explain the definition of equation (2.1.18). Our primary sources for stochastic differential geometry are [43, 47].

Notation . We use bold font to denote $\mathcal{M}$-valued processes and vector fields on $\mathcal{M}$. Coordinates of these processes are written in the standard font. For example, let $\varphi = (\varphi^1, ..., \varphi^d)$ be local coordinates on an open set $U \subset \mathcal{M}$. The i-th coordinate of an $\mathcal{M}$-valued process $\boldsymbol{X}_t$ is denoted by $X_t^i = \varphi^i(\boldsymbol{X}_t)$. A vector field $\boldsymbol{V}$ on $\mathcal{M}$ can be described in coordinates as $\boldsymbol{V}(X) = V^i(X)\partial_i = V^i(X)\frac{\partial}{\partial \varphi^i}$, $X \in U$. For a function $f \in C^\infty(\mathcal{M})$, $\boldsymbol{V}(f) = \boldsymbol{V}(\mathrm{d}f) = \mathrm{d}f(\boldsymbol{V})$ gives the directional derivative of f along $\boldsymbol{V}$.

2.2.2 SDE, martingales on manifolds

The Stratonovich formulation is used to define SDE on manifolds. A typical SDE on $\mathcal{M}$ is as follows. Assume given l vector fields $\{\boldsymbol{V}_1, ..., \boldsymbol{V}_l\}$ on $\mathcal{M}$, an $\mathbb{R}^l$-valued driving semimartingale Z, and write

$$\mathrm{d}\boldsymbol{X}_t = \sum_{\alpha=1}^{l} \boldsymbol{V}_\alpha(\boldsymbol{X}_t) \circ \mathrm{d}Z_t^\alpha. \tag{2.2.1}$$

The use of the Stratonovich formulation ensures that the usual chain rule applies, ensuring coordinate invariance of the solution to the SDE. The solution theory uses semimartingales in the following way.

Definition 1. Let $\mathcal{M}$ be a smooth manifold and $(\Omega, \mathcal{F}_*, \mathbb{P})$ a filtered probability space. Let τ be an $\mathcal{F}_*$-stopping time. A continuous $\mathcal{M}$-valued process $\boldsymbol{X}_t$ defined on $[0, \tau)$ is called an $\mathcal{M}$-valued semimartingale if $f(\boldsymbol{X}_t)$ is a real valued semimartingale on $[0, \tau)$ for all $f \in C^\infty(\mathcal{M})$.

Definition 2. An $\mathcal{M}$-valued semimartingale $\boldsymbol{X}$ defined up to a stopping time τ is a solution to the SDE (2.2.1) if for all $f \in C^\infty(\mathcal{M})$,

$$f(\boldsymbol{X}_t) = f(\boldsymbol{X}_0) + \sum_{\alpha=1}^{l} \int_0^t \boldsymbol{V}_\alpha f(\boldsymbol{X}_t) \circ \mathrm{d}Z_s^\alpha, \quad 0 \le t < \tau. \tag{2.2.2}$$

These two definitions are intrinsic. Only the differentiable structure on $\mathcal{M}$ has been used to define the SDE (2.2.1), since we have only used the concept of vector fields. However, in order to define martingales taking values in $\mathcal{M}$, the manifold must be equipped with a connection ∇. For Riemannian manifolds, we always use the Levi-Civita connection.

Definition 3. Assume $\mathcal{M}$ is a smooth manifold equipped with a connection ∇. An $\mathcal{M}$-valued semimartingale $\boldsymbol{X}$ is called a ∇-martingale if and only if the following process is a $\mathbb{R}$-valued local martingale for all $f \in C^\infty(\mathcal{M})$:

$$N^f(\boldsymbol{X})_t := f(\boldsymbol{X}_t) - f(\boldsymbol{X}_0) - \frac{1}{2}\int_0^t \nabla^2 f(\mathrm{d}\boldsymbol{X}_s, \mathrm{d}\boldsymbol{X}_s), \tag{2.2.3}$$

where $\nabla^2 f$ is the Hessian of f under ∇.

In our model, the driving semimartingales are Wiener processes or just t. The SDE we are concerned with have the following form:

$$\mathrm{d}\boldsymbol{X}_t = \sum_{\alpha=1}^{l} \Big(\boldsymbol{V}_\alpha(\boldsymbol{X}_t) \circ \mathrm{d}W_t^\alpha - \frac{1}{2}\nabla_{\boldsymbol{V}_\alpha}\boldsymbol{V}_\alpha(\boldsymbol{X}_t)\mathrm{d}t \Big) \tag{2.2.4}$$

where $(W_t^1, ..., W_t^l)$ are independent Wiener processes, and $\nabla_{\boldsymbol{V}_\alpha}\boldsymbol{V}_\alpha$ is the covariant derivative of $\boldsymbol{V}_\alpha$ along itself. The deterministic correction is introduced in equation (2.2.4) in order that we could formally regard equation (2.2.4) as the Itô SDE

$$\mathrm{d}\boldsymbol{X}_t = \sum_{\alpha=1}^{l} \boldsymbol{V}_\alpha(\boldsymbol{X}_t)\, \mathrm{d}W_t^\alpha. \tag{2.2.5}$$

More precisely, we have

Proposition 1. *The solution to the SDE* (2.2.4) *is a* ∇*-martingale.*

Proof. Inserting equation (2.2.2) into equation (2.2.3), we only need to show that the following process is a local martingale [43, Prop 2.5.2,pp. 56] :

$$N^f(\boldsymbol{X})_t = \sum_{\alpha=1}^{l} \Big(\int_0^t \Big(\boldsymbol{V}_\alpha f(\boldsymbol{X}_t) \circ \mathrm{d}W_s^\alpha - \frac{1}{2}(\nabla_{\boldsymbol{V}_\alpha}\boldsymbol{V}_\alpha) f(\boldsymbol{X}_t) \Big) \mathrm{d}t - \frac{1}{2}\int_0^t \nabla^2 f(\boldsymbol{V}_\alpha, \boldsymbol{V}_\alpha)(\boldsymbol{X}_t)\mathrm{d}t \Big). \tag{2.2.6}$$

But since $\boldsymbol{V}(\boldsymbol{V}f) = (\nabla_{\boldsymbol{V}}\boldsymbol{V})f + \nabla_{\boldsymbol{V}}(\mathrm{d}f)(\boldsymbol{V}) = (\nabla_{\boldsymbol{V}}\boldsymbol{V})f + \nabla^2 f(\boldsymbol{V}, \boldsymbol{V})$, we combine the last two terms to obtain

$$\begin{aligned} N^f(\boldsymbol{X})_t &= \sum_{\alpha=1}^{l} \Big(\int_0^t \boldsymbol{V}_\alpha f(\boldsymbol{X}_t) \circ \mathrm{d}W_s^\alpha - \frac{1}{2}\int_0^t \boldsymbol{V}_\alpha(\boldsymbol{V}_\alpha f)(\boldsymbol{X}_t)\mathrm{d}t \Big) \\ &= \sum_{\alpha=1}^{l} \int_0^t \boldsymbol{V}_\alpha f(\boldsymbol{X}_t)\mathrm{d}W_t^\alpha. \end{aligned} \tag{2.2.7}$$

□

2.2.3 Brownian motion and orthogonal projections

When the manifold is Riemannian we may characterize Brownian motion in several ways. Orthonormal bases provide a direct approach.

Corollary 1. *Let* $\{\boldsymbol{V}_\alpha\}_{\alpha=1}^n$ *be an orthonormal basis for* $T\mathcal{M}$*. Then the solution to* *the SDE*

$$\mathrm{d}\boldsymbol{X}_t = \sum_{\alpha=1}^{n}(\boldsymbol{V}_\alpha \circ \mathrm{d}W_t^\alpha - \frac{1}{2}\nabla_{\boldsymbol{V}_\alpha}\boldsymbol{V}_\alpha \mathrm{d}t) \tag{2.2.8}$$

has generator $\Delta = \mathrm{Tr}(\nabla^2)$*, so that it is a Brownian motion on* $\mathcal{M}$*.*

Proof. For any $f \in C^\infty(\mathcal{M})$, we have

$$\begin{aligned} f(\boldsymbol{X}_t) - f(\boldsymbol{X}_0) =& \sum_{\alpha=1}^{n}\int_0^t \Big(\boldsymbol{V}_\alpha f(\boldsymbol{X}_s) \circ \mathrm{d}W_s^\alpha - \frac{1}{2}\nabla_{\boldsymbol{V}_\alpha}\boldsymbol{V}_\alpha f(\boldsymbol{X}_s)\mathrm{d}s\Big) \\ =& \sum_{\alpha=1}^{n}\int_0^t \frac{1}{2}\nabla^2 f(\boldsymbol{V}_\alpha, \boldsymbol{V}_\alpha)(\boldsymbol{X}_s)\mathrm{d}s \\ =& \int_0^t \frac{1}{2}\Delta f(\boldsymbol{X}_s)\mathrm{d}s\,. \end{aligned} \tag{2.2.9}$$

□

At this stage we are able to define the projection operator P onto a submanifold $\mathcal{O} \subset \mathcal{M}$ clearly. For drift terms in an SDE, no new ideas are required: we simply project the drift vector to $T\mathcal{O}$. For diffusion term that is an ∇-martingale in an SDE, e.g. with a single driving vector field,

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{V} \circ \mathrm{d}W_t - \frac{1}{2}\nabla_{\boldsymbol{V}}\boldsymbol{V}\mathrm{d}t, \tag{2.2.10}$$

we define

$$P(\mathrm{d}\boldsymbol{X}_t) = P(\boldsymbol{V}) \circ \mathrm{d}W_t - \frac{1}{2}\nabla_{P(\boldsymbol{V})}P(\boldsymbol{V})\mathrm{d}t\,. \tag{2.2.11}$$

To understand this projection, we have the following Lemma:

Lemma 4. *Let* $f \in C^\infty(\mathcal{M})$*,* $\boldsymbol{X}_t$ *solves SDE* (2.2.10) *and is a* ∇*-martingale. By*

(2.2.3), *we could write*

$$\mathrm{d}f(\boldsymbol{X}_t) = \frac{1}{2}\nabla^2 f(\boldsymbol{V}, \boldsymbol{V})\mathrm{d}t + \text{ \textit{some local martingale}}. \tag{2.2.12}$$

Let $\boldsymbol{X}_t^p$ *be the solution to the SDE:*

$$\mathrm{d}\boldsymbol{X}_t^p = P(\boldsymbol{V}) \circ \mathrm{d}W_t - \frac{1}{2}\nabla_{P(\boldsymbol{V}} P(\boldsymbol{V})\mathrm{d}t\,. \tag{2.2.13}$$

then we have

$$\mathrm{d}f(\boldsymbol{X}_t^p) = \frac{1}{2}\nabla^2 f(P(\boldsymbol{V}), P(\boldsymbol{V}))\mathrm{d}t + \text{ \textit{some local martingale}}. \tag{2.2.14}$$

For an SDE with both diffusion and drift terms, we decompose the RHS of the SDE into the sum of a ∇-martingale diffusion term and drift term and project them respectively. The result is that the evolution of $f(X_t^p)$, for $f \in C^\infty(\mathcal{M})$ and a solution to the projected SDE $\boldsymbol{X}_t^p$, only contains the values of ∇f and $\nabla^2 f$ in the subspace to be projected.

When we have an orthonormal basis $\{\boldsymbol{V}_\alpha\}_{\alpha=1}^n$ of $T\mathcal{M}$ such that $\{\boldsymbol{V}_\alpha\}_{\alpha=1}^d$ is also an orthonormal basis of $T\mathcal{O}$, and $\boldsymbol{X}_t$ satisfies equation (2.2.8), we have

$$P(\mathrm{d}\boldsymbol{X}_t) = \sum_{\alpha=1}^{d}(\boldsymbol{V}_\alpha \circ \mathrm{d}W_t^\alpha - \frac{1}{2}\nabla_{\boldsymbol{V}_\alpha}\boldsymbol{V}_\alpha \mathrm{d}t), \tag{2.2.15}$$

$$P^\perp(\mathrm{d}\boldsymbol{X}_t) = \sum_{\alpha=d+1}^{d}(\boldsymbol{V}_\alpha \circ \mathrm{d}W_t^\alpha - \frac{1}{2}\nabla_{\boldsymbol{V}_\alpha}\boldsymbol{V}_\alpha \mathrm{d}t). \tag{2.2.16}$$

This is the form in which we will prove Theorem 3 and Theorem 4.

2.2.4 Some matrix algebra to be used

In this section, we introduce some matrix algebra to be frequently used in the following sections.

We first introduce a matrix representation of $\mathfrak{u}(n)$, the Lie algebra of unitary group $\mathrm{U}(n)$. Define

$$\alpha_{k,l} := \frac{e_l e_k^T - e_k e_l^T}{\sqrt{2}}\,, \qquad 1 \le k < l \le n\,, \tag{2.2.17}$$

$$\beta_k := \boldsymbol{i} e_k e_k\,, \qquad 1 \le k \le n\,, \tag{2.2.18}$$

$$\beta_{k,l} := \boldsymbol{i}\frac{e_l e_k^T + e_k e_l^T}{\sqrt{2}}\,, \qquad 1 \le k < l \le n\,. \tag{2.2.19}$$

then $\mathfrak{u}(n)$ is the linear space spanned by these vectors

$$\begin{aligned}\mathfrak{u}(n) &= \mathrm{Span}(\{\alpha_{k,l}\}_{1\le k<l\le n}^n \cup \{\beta_k\}_{1\le k\le n} \cup \{\beta_{k,l}\}_{1\le k<l\le n})\,, \\ &= \left\{A \in M_{n\times n} | A = -A^\dagger := -(\bar{A})^T\right\}.\end{aligned} \tag{2.2.20}$$

The validity of this statement can be verified by dimension counting.

We should note that this basis of $\mathfrak{u}(n)$ is also orthonormal under the inner product $\mathrm{Re}(\mathrm{Tr}(UV^\dagger))$:

$$\begin{aligned}&\mathrm{Re}(\mathrm{Tr}(\alpha_{i,j}\alpha_{k,l}^\dagger)) = \delta_{ik}\delta_{jl}\,, \quad \mathrm{Re}(\mathrm{Tr}(\beta_i\beta_k^\dagger)) = \delta_{ik}\,, \quad \mathrm{Re}(\mathrm{Tr}(\beta_{i,j}\beta_{k,l}^\dagger)) = \delta_{ik}\delta_{jl}\,, \\ &\mathrm{Re}(\mathrm{Tr}(\alpha_{i,j}\beta_k^\dagger)) = 0\,, \qquad \mathrm{Re}(\mathrm{Tr}(\alpha_{i,j}\beta_{k,l}^\dagger)) = 0\,, \quad \mathrm{Re}(\mathrm{Tr}(\beta_i\beta_{k,l}^\dagger)) = 0\,.\end{aligned} \tag{2.2.21}$$

Their commutator is defined to be $[A,B] := AB - BA$. We are not going to write down all commutators, but the following property will be used:

$$\mathrm{Tr}([A,B]B^\dagger) = -\mathrm{Tr}([A,B]B) = -\mathrm{Tr}(AB^2 - BAB) = 0\,. \tag{2.2.22}$$

which means that the commutator $[A, B]$ is perpendicular to B (and A by symmetry) under inner product. By the completeness of the inner product $\mathrm{Re}(\mathrm{Tr}(UV^\dagger))$, we have that the commutator $[A, B]$, for A, B chosen from the basis above, is a linear combination of basis vectors other than A and B.

In the analysis of symmetric matrix $X \in \mathrm{Sym}(n)$, its diagonalization $X = PMP^T$ always exists, where $P \in \mathrm{O}(n)$ and M is $n \times n$ diagonal matrix. For complex symmetric matrices we also have the following decomposition.

Lemma 5 (Takagi, [87, Theorem 2])**.** *Each complex symmetric matrix A can be decomposed to $A = QMQ^T$, where $M = \mathrm{diag}\,(\mu_1, ..., \mu_n)$ is a real nonnegative diagonal matrix and $Q \in \mathrm{U}(n)$.*

Such decomposition provides an isomorphism between linear vector spaces: the tangent spaces $T\mathrm{Sym}(n, \mathbb{C})$ and $T\mathbb{R}^n \otimes \mathfrak{u}(n)$. To be specific, we consider the following basis of $T\mathrm{Sym}(n, \mathbb{C})$: define E^R, E^I matrices as:

$$E^R_k = e_k e_k^T\,, \qquad E^I_k = \boldsymbol{i} e_k e_k^T\,, \qquad 1 \le k \le n\,, \tag{2.2.23}$$

$$E^I_{k,l} = \boldsymbol{i}\frac{e_k e_l^T + e_l e_k^T}{\sqrt{2}}\,, \qquad E^R_{k,l} = \frac{e_k e_l^T + e_l e_k^T}{\sqrt{2}}\,, \qquad 1 \le k < l \le n\,. \tag{2.2.24}$$

and define tangent vector fields:

$$\boldsymbol{l}_k : \qquad \boldsymbol{l}_k(\mathrm{d}X) = QE^R_k Q^T\,, \qquad 1 \le k \le n\,, \tag{2.2.25}$$

$$\boldsymbol{u}_{(k)} : \qquad \boldsymbol{u}_{(k)}(\mathrm{d}X) = 2\mu_k QE^I_k Q^T\,, \qquad 1 \le k \le n\,, \tag{2.2.26}$$

$$\boldsymbol{u}_{(k,l,1)} : \qquad \boldsymbol{u}_{(k,l,1)}(\mathrm{d}X) = (\mu_k + \mu_l) QE^I_{k,l} Q^T\,, \qquad 1 \le k < l \le n\,, \tag{2.2.27}$$

$$\boldsymbol{u}_{(k,l,2)} : \qquad \boldsymbol{u}_{(k,l,2)}(\mathrm{d}X) = (\mu_k - \mu_l) QE^R_{k,l} Q^T\,, \qquad 1 \le k < l \le n\,. \tag{2.2.28}$$

We also use the following basis of $T\mathbb{R}^n \otimes \mathfrak{u}(n)$:

$$\boldsymbol{l}'_k: \qquad \boldsymbol{l}'_k(\mathrm{d}M) = e_k e_k^T\,, \qquad \boldsymbol{l}'_k(\mathrm{d}Q) = 0\,, \qquad 1 \le k \le n\,, \tag{2.2.29}$$

$$\boldsymbol{u}'_{(k)}: \qquad \boldsymbol{u}'_{(k)}(\mathrm{d}M) = 0\,, \qquad \boldsymbol{u}'_{(k)}(\mathrm{d}Q) = Q\beta_k\,, \qquad 1 \le k \le n\,, \tag{2.2.30}$$

$$\boldsymbol{u}'_{(k,l,1)}: \qquad \boldsymbol{u}'_{(k,l,1)}(\mathrm{d}M) = 0\,, \qquad \boldsymbol{u}'_{(k,l,1)}(\mathrm{d}Q) = Q\beta_{k,l}\,, \qquad 1 \le k < l \le n\,, \tag{2.2.31}$$

$$\boldsymbol{u}'_{(k,l,2)}: \qquad \boldsymbol{u}'_{(k,l,2)}(\mathrm{d}M) = 0\,, \qquad \boldsymbol{u}'_{(k,l,2)}(\mathrm{d}Q) = Q\alpha_{k,l}\,, \qquad 1 \le k < l \le n\,. \tag{2.2.32}$$

We claim that these two set of vector fields are exactly the same vector fields. More precisely, let $\pi : \mathbb{R}^n \otimes \mathrm{U}(n) \to \mathrm{Sym}(n, \mathbb{C}), \pi(M, Q) = QMQ^T$ be the bijection between $\mathbb{R}^n \otimes \mathrm{U}(n)$ and $\mathrm{Sym}(n, \mathbb{C})$, and $\pi_* : T\mathbb{R}^n \otimes \mathfrak{u}(n) \to T\mathrm{Sym}(n, \mathbb{C})$ be the differential of π, or the pushforward map, we have $\pi_*(\boldsymbol{v}') = \boldsymbol{v}$ for $\boldsymbol{v}$ being any vector fields defined above.

To see this, one should check the following matrix identities:

$$[e_k e_k^T, M] = 0\,, \qquad [\alpha_{k,l}, M] = (\mu_k - \mu_l)E^R_{k,l}\,,$$
$$\{\beta_k, M\} = 2\mu_k E^I_k\,, \qquad \{\beta_{k,l}, M\} = (\mu_k + \mu_l)E^I_{k,l}\,.$$

for M being diagonal matrix, where $\{A, B\} = AB + BA$ is the anti-commutator. Then by taking the differential of $X = QMQ^T$:

$$\begin{aligned} \mathrm{d}X =& Q(\mathrm{d}M + Q^{-1}\mathrm{d}QM + M\mathrm{d}Q^T(Q^T)^{-1})Q^T \\ (\mathrm{d}\omega := Q^{-1}\mathrm{d}Q) =& Q(\mathrm{d}M + \mathrm{d}\omega M + M\mathrm{d}\omega^T)Q^T \\ =& Q(\mathrm{d}M + \mathrm{d}\omega M - M\mathrm{d}\bar{\omega})Q^T \\ =& Q(\mathrm{d}M + [\mathrm{Re}(\mathrm{d}\omega), M] + \{\boldsymbol{i}\,\mathrm{Im}(\mathrm{d}\omega), M\})Q^T\,, \end{aligned} \tag{2.2.33}$$

we may verify for example that

$$\begin{aligned}\boldsymbol{u}'_{(k,l,1)}(\mathrm{d}X) =& Q\{\beta_{k,l}, M\}Q^T = Q(\mu_k+\mu_l)(\frac{e_k e_l^T + e_l e_k^T}{\sqrt{2}})Q^T \\ =& (\mu_k+\mu_l)QE^I_{k,l}Q^T = \boldsymbol{u}_{(k,l,1)}(\mathrm{d}X)\,.\end{aligned} \tag{2.2.34}$$

2.3 Proof of Theorem 3

The proof of Theorem 3 consists of two parts. First, we express the SDE (2.1.18) for $\boldsymbol{Z}_t$ on $\mathcal{H}$ in a more explicit form using the orthonormal bases described in Section 2.2. Next, we use eigenvalue perturbation theory, the definition of the cross-ratio $\mathfrak{R}(Z_t, iI_n)$ and the change of variables from λ_t to σ_t to show that σ_t satisfies (2.1.7). Almost all the work lies in explicit computations related to the change of basis. The eigenvalue perturbation theory is standard, except near repeated eigenvalues. The well-posedness theory in Section 2.5 allows us to focus on the case of distinct eigenvalues.

The outline of the computation of the basis is as follows. Following Siegel, we first change variables from $\mathcal{H}$ to the equivalent of the unit disk for complex symmetric matrices (the domain $\mathcal{D}$ below). This is a convenient coordinate system, since it makes several properties of the cross-ratio matrix obvious. We will find an orthonormal basis $\{\boldsymbol{L}_k\}_{k=1}^n \cup \{\boldsymbol{U}_i\}_{i=1}^{n^2}$ for $T\mathcal{H}$, such that $\{\boldsymbol{U}_i\}_{i=1}^{n^2}$ is also an orthonormal basis of $T\mathcal{O}_\lambda$, and $\{\boldsymbol{L}_k\}_{k=1}^n \subset T\mathcal{O}_\lambda^\perp$. By Riemannian submersion, $\{\boldsymbol{L}_k\}_{k=1}$ also gives an orthonormal basis of $T\Lambda$ (identifying $\boldsymbol{L}_k$ with $\mathrm{d}\pi(\boldsymbol{L}_k)$). Then by Corollary 1, we are able to construct Brownian motion on $(\mathcal{H}, g)$; equations (2.2.15) and (2.2.16) then complete the definition of the SDE (2.1.7).

2.3.1 The unit disk $\mathcal{D}$

The domain of symmetric complex matrices defined by

$$\mathcal{D} = \{W = W^T \,\big|\, I - W\bar{W} \succ 0\} \tag{2.3.1}$$

is the analogue of the open unit disk in $\mathbb{C}$. Since $W = W^T$, the matrix $I - W\bar{W}$ is Hermitian and the notation $I - W\bar{W} \succ 0$ means that it is positive definite.

The conformal mapping $z \mapsto \frac{z-i}{z+i}$ maps the Poincaré upper half-plane to the unit disk in $\mathbb{C}$. We write the analogous transformation for the Siegel half-space $\mathcal{H}$ as

$$R = \varphi(Z) = (Z - \boldsymbol{i}I_n)(Z + \boldsymbol{i}I_n)^{-1}. \tag{2.3.2}$$

Since $Z = Z^T$ and these matrices commute, we have $R = R^T$ for $Z \in \mathcal{H}$, a property that we will use repeatedly. Siegel showed that $\mathcal{D} = \varphi(\mathcal{H})$ by adapting the proof of conformal equivalence between the upper half-plane and the unit disk to several complex variables (see [83, II.4]).

The transformation (2.3.2) plays an important role in our work because the cross-ratio matrix $\mathfrak{R}(Z, \boldsymbol{i}I_n)$ may be written

$$\mathfrak{R}(Z, \boldsymbol{i}I_n) = R\bar{R}. \tag{2.3.3}$$

Since $R = R^T$ we find immediately that

$$\bar{\mathfrak{R}}^T = (\bar{R}R)^T = R^T\bar{R}^T = R\bar{R} = \mathfrak{R}. \tag{2.3.4}$$

Thus, $\mathfrak{R}$ is Hermitian positive semi-definite. This property allows us to obtain precise descriptions of the orbits $\mathcal{O}_\lambda$ and the metric $g_{\mathcal{H}}$ as outlined below. We begin with a classical

Lemma 6. *The domain $\mathcal{D}$ may be expressed as*

$$\mathcal{D} = \varphi(\mathcal{H}) = \left\{QMQ^T \middle| Q \in U(n), M \text{ is diagonal}, \ 0 \le \mu_i < 1, i = 1, ..., n\right\}. \tag{2.3.5}$$

Proof. The first equality in equation (2.3.5) follows from [83, II.4]. The second equality is also straightforward. We decompose $R = QMQ^T$ using Lemma 5 and invert equation (2.3.2) to find

$$\begin{aligned}\mathrm{Im}(Z) =&(I_n - R)^{-1}(I_n - R\bar{R})(I_n - \bar{R})^{-1} \\ =&(\bar{Q}^T - MQ^T)^{-1}(I_n - M^2)(Q - \bar{Q}M)^{-1}.\end{aligned} \tag{2.3.6}$$

It is immediate that $\mathrm{Im}(Z) \succ 0$ if and only if $\mu_i^2 < 1$ for $i = 1, \ldots, n$. □

Lemma 7. *For $Z \in \mathcal{H}$ and $\lambda = \operatorname{eig} \mathfrak{R}(Z, \boldsymbol{i}I_n)$, the orbit $\mathcal{O}_\lambda$ is represented by*

$$\mathcal{O}_\lambda = \varphi^{-1}\Big(\{QMQ^T | M = \operatorname{diag}(\mu_1, ..., \mu_n), Q \in U(n)\}\Big) \tag{2.3.7}$$

where $\lambda = (\mu_1^2, ..., \mu_n^2)$.

Proof. By equation (2.3.3) and Lemma 5, the eigenvalues of $\mathfrak{R}(Z, \boldsymbol{i}I_n)$ are just the diagonal elements $\{\mu_i^2\}_{i=1}^n$. Thus,

$$\varphi(\mathcal{O}_\lambda) = \{QMQ^T | M = \operatorname{diag}(\mu_1, ..., \mu_n), Q \in U(n)\} \tag{2.3.8}$$

where $\lambda = (\mu_1^2, ..., \mu_n^2)$. □

2.3.2 The orthonormal basis

Now we will construct an orthonormal basis of $T\mathcal{H}$ and $T\mathcal{O}_\lambda$. As $\varphi : \mathcal{H} \to \mathcal{D}$ is an analytic bijection and we will equip $\mathcal{D}$ with metric that is compatible with φ later, we can regard $(\mathcal{H}, g_{\mathcal{H}})$ and $(\mathcal{D}, g_{\mathcal{D}})$ as the same Riemannian manifold, while Z and R are just two coordinate systems on the this manifold. Also, vector fields related by the pushforward map $\mathrm{d}\pi$ (e.g. $\boldsymbol{u}$ in (2.2.25) and $\boldsymbol{u}'$ in (2.2.29)) are regarded as representations of the same vector field under different coordinate systems.

In this section we will mainly use R and its diagonalization $R = QMQ^T$ as coordinate system. According to section 2.2.4, we already have a basis (2.2.25) of $T_R\mathcal{D} = T_R\mathrm{Sym}(n,\mathbb{C})$ for $R \in \mathcal{D}$.

By definition of $\boldsymbol{u}$ vector fields (2.2.29), we also know that for each $Z \in \mathcal{O}_\lambda$

$$T_Z\mathcal{O}_\lambda = \mathrm{Span}\big(\{\boldsymbol{u}_{(k)}\}_{k=1}^n \cup \{\boldsymbol{u}_{(k,l,1)}\}_{1\le k<l\le n} \cup \{\boldsymbol{u}_{(k,l,2)}\}_{1\le k<l\le n}\big). \tag{2.3.9}$$

Now we will compute the metric $g_\mathcal{D}$ that is compatible with settings above. In view of bilinear form, we just need to compute the transformation coefficient of $\mathrm{d}\pi$ and get $g_\mathcal{D}$. In a geometric point of view, we have $g_\mathcal{D} = g_\mathcal{H}$ in the sense that they are the same tensor in $\mathfrak{X}(T^*\mathcal{H} \otimes T^*\mathcal{H})$ represented by different coordinate systems. Under the geometric point of view, we use

$$\begin{aligned} R =&(Z - \boldsymbol{i}I_n)(Z + \boldsymbol{i}I_n)^{-1}, \quad Z = \boldsymbol{i}(I_n + R)(I_n - R)^{-1}\,, &(2.3.10)\\ \mathrm{d}Z =&2\boldsymbol{i}(I_n - R)^{-1}\mathrm{d}R(I_n - R)^{-1}\,,\\ \text{and } Y =&\frac{1}{2\boldsymbol{i}}(Z - \bar{Z}) = (I_n - R)^{-1}(I_n - R\bar{R})(I_n - \bar{R})^{-1} &(2.3.11)\\ =&\bar{Y} = (I_n - \bar{R})^{-1}(I_n - \bar{R}R)(I_n - R)^{-1}\,, \end{aligned}$$

and compute

$$\begin{aligned} g_\mathcal{H} =&\mathrm{Tr}(Y^{-1}\mathrm{d}ZY^{-1}\mathrm{d}\bar{Z}) = \mathrm{Tr}(Y^{-1}\mathrm{d}Z(\bar{Y})^{-1}\mathrm{d}\bar{Z}) &(2.3.12)\\ =&\mathrm{Tr}\Big([(I_n - \bar{R})(I_n - R\bar{R})^{-1}(I_n - R)][2\boldsymbol{i}(I_n - R)^{-1}\mathrm{d}R(I_n - R)^{-1}]\\ &[(I_n - R)(I_n - \bar{R}R)^{-1}(I_n - \bar{R})][(-2\boldsymbol{i})(I_n - \bar{R})^{-1}\mathrm{d}\bar{R}(I_n - \bar{R})^{-1}]\Big)\\ =&4\,\mathrm{Tr}\Big((I_n - R\bar{R})^{-1}\mathrm{d}R(I_n - \bar{R}R)^{-1}\mathrm{d}\bar{R}\Big) = g_\mathcal{D}. \end{aligned}$$

Notice that when $n = 1$, $g_\mathcal{D} = \frac{4\mathrm{d}z\mathrm{d}\bar{z}}{(1-|z|^2)^2}$ coincides with the Poincaré metric on the unit disk. From now on we will only use $g_\mathcal{H}$ for both $g_\mathcal{H}$ and $g_\mathcal{D}$, as they are conceptually the same geometric object. In the following theorem, we will show that the metric $g_{\mathcal{H}}$ is diagonalized under the basis (2.2.25). Thus, if we normalize the basis (2.2.25) we will get an orthonormal basis.

Theorem 9. *The following vector fields* $\{\boldsymbol{L}_k\}_{k=1}^n \cup \{\boldsymbol{U}_{(k)}\}_{k=1}^n \cup \{\boldsymbol{U}_{(k,l,m)}\}_{1\le k<l\le n, m=1,2}$ *give an orthonormal basis of* $\mathfrak{X}(T\mathcal{H})$*. Moreover,* $\{\boldsymbol{U}_{(k)}(Z)\}_{k=1}^n \cup \{\boldsymbol{U}_{(k,l,m)}(Z)\}_{1\le k<l\le n, m=1,2}$ *give an orthonormal basis of* $T_Z\mathcal{O}_\lambda$ *for each* $Z \in \mathcal{O}_\lambda$ *and* $\lambda \in \mathcal{W}$*.*

At a point $\varphi(Z) = R = QMQ^T$*, where* $M = \operatorname{diag}(\tanh\frac{\sigma^1}{2}, ..., \tanh\frac{\sigma^n}{2})$*, the vector fields are defined as follows*

$$\boldsymbol{L}_k(\mathrm{d}R) = \frac{1}{2\cosh^2\frac{\sigma^k}{2}} Q\Big(e_k e_k^T\Big)Q^T, \qquad 1 \le k \le n, \tag{2.3.13}$$

$$\boldsymbol{U}_{(k)}(\mathrm{d}R) = \frac{1}{2\cosh^2\frac{\sigma^k}{2}} Q\Big(\boldsymbol{i} e_k e_k^T\Big)Q^T, \qquad 1 \le k \le n, \tag{2.3.14}$$

$$\boldsymbol{U}_{(k,l,1)}(\mathrm{d}R) = \frac{1}{2\cosh\frac{\sigma^k}{2}\cosh\frac{\sigma^l}{2}} Q\Big(\frac{\boldsymbol{i}(e_k e_l^T + e_l e_k^T)}{\sqrt{2}}\Big)Q^T, \qquad 1 \le k < l \le n, \tag{2.3.15}$$

$$\boldsymbol{U}_{(k,l,2)}(\mathrm{d}R) = \frac{1}{2\cosh\frac{\sigma^k}{2}\cosh\frac{\sigma^l}{2}} Q\Big(\frac{e_k e_l^T + e_l e_k^T}{\sqrt{2}}\Big)Q^T, \qquad 1 \le k < l \le n. \tag{2.3.16}$$

In order to simplify notation, we will use $\{\boldsymbol{U}_i\}_{i=1}^{n^2}$ to denote the union $\{\boldsymbol{U}_{(k)}\}_{k=1}^n \cup \{\boldsymbol{U}_{(k,l,m)}\}_{1\le k<l\le n, m=1,2}$ and use $\boldsymbol{U}_i$ for any vector field in $\{\boldsymbol{U}_i\}_{i=1}^{n^2}$, when there is no confusion.

Proof. Notice that all $\boldsymbol{L}$ and $\boldsymbol{U}$ vectors are just normalized $\boldsymbol{l}$ and $\boldsymbol{u}$ vectors. We only need to prove that $g_{\mathcal{H}}$ is diagonalized under $\boldsymbol{l}$ and $\boldsymbol{u}$ basis (2.2.25), and extract all coefficient of squares.

To do this, we first transform $g_{\mathcal{H}}$ into (M, Q) system. Use (2.2.33), we get $g_{\mathcal{H}}$

$$
\begin{aligned}
g_{\mathcal{H}} =& 4\operatorname{Tr}\Big((I_n - R\bar{R})^{-1}\mathrm{d}R(I_n - \bar{R}R)^{-1}\mathrm{d}\bar{R}\Big) \qquad (2.3.17)\\
=& 4\operatorname{Tr}\Big((I_n - M^2)^{-1}(\mathrm{d}M + [\operatorname{Re}(\mathrm{d}\omega), M] + \{\boldsymbol{i}\operatorname{Im}(\mathrm{d}\omega), M\})\\
&\quad (I_n - M^2)^{-1}(\mathrm{d}M + [\operatorname{Re}(\mathrm{d}\omega), M] - \{\boldsymbol{i}\operatorname{Im}(\mathrm{d}\omega), M\})\Big)\\
=& 4\Big(\operatorname{Tr}((I_n - M^2)^{-1}(\mathrm{d}M + [\operatorname{Re}(\mathrm{d}\omega), M])(I_n - M^2)^{-1}(\mathrm{d}M + [\operatorname{Re}(\mathrm{d}\omega), M]))\\
&\quad + \operatorname{Tr}((I_n - M^2)^{-1}\{\operatorname{Im}(\mathrm{d}\omega), M\}(I_n - M^2)^{-1}\{\operatorname{Im}(\mathrm{d}\omega), M\})\Big)\\
=& 4\Big(\operatorname{Tr}((I_n - M^2)^{-2}\mathrm{d}M\mathrm{d}M)\\
&\quad + \operatorname{Tr}((I_n - M^2)^{-1}[\operatorname{Re}(\mathrm{d}\omega), M](I_n - M^2)^{-1}[\operatorname{Re}(\mathrm{d}\omega), M])\\
&\quad + \operatorname{Tr}((I_n - M^2)^{-1}\{\operatorname{Im}(\mathrm{d}\omega), M\}(I_n - M^2)^{-1}\{\operatorname{Im}(\mathrm{d}\omega), M\})\Big).
\end{aligned}
$$

where we used the fact that $(I_n - M^2)$ and $\mathrm{d}M$ are diagonal and $[\operatorname{Re}(\mathrm{d}\omega), M]$ has zero diagonal.

Take linear combination of basis with constant coefficients

$$
\boldsymbol{v} = \sum_{k=1}^{n} c_k \boldsymbol{l}_k + \sum_{k=1}^{n} d_k \boldsymbol{u}_{(k)} + \sum_{1\le k<l\le n} d_{k,l,1}\boldsymbol{u}_{(k,l,1)} + \sum_{1\le k<l\le n} d_{k,l,2}\boldsymbol{u}_{(k,l,2)}, \qquad (2.3.18)
$$

we find that $g_{\mathcal{H}}(\boldsymbol{v}, \boldsymbol{v})$ becomes a sum of squares

$$
\begin{aligned}
g_{\mathcal{H}}(\boldsymbol{v}, \boldsymbol{v}) =& \sum_{k=1}^{n} \frac{4}{(1-\mu_k^2)^2} c_k^2 + \sum_{k=1}^{n} \frac{4(2\mu_k)^2}{(1-\mu_k^2)^2} d_k^2 \qquad (2.3.19)\\
&+ \sum_{1\le k<l\le n} \frac{4(\mu_k + \mu_l)^2}{(1-\mu_k^2)(1-\mu_l^2)} d_{k,l,1}^2 + \sum_{1\le k<l\le n} \frac{4(\mu_k - \mu_l)^2}{(1-\mu_k^2)(1-\mu_l^2)} d_{k,l,2}^2.
\end{aligned}
$$

Thus, the vectors $\{\boldsymbol{l}_k\}_{k=1}^n \cup \{\boldsymbol{u}_{(k)}\}_{k=1}^n \cup \{\boldsymbol{u}_{(k,l,1)}\}_{1\le k<l\le n} \cup \{\boldsymbol{u}_{(k,l,2)}\}_{1\le k<l\le n}$ give an orthogonal basis. We normalize this basis to obtain an orthonormal basis (2.3.13).

For example, we can check that

$$
\begin{aligned}
g_{\mathcal{H}}(\boldsymbol{L}_k, \boldsymbol{L}_k) =& g_{\mathcal{H}}(\frac{1}{2\cosh\frac{\sigma^k}{2}}\boldsymbol{l}_k, \frac{1}{2\cosh\frac{\sigma^k}{2}}\boldsymbol{l}_k) \\
=& \frac{4}{(1-\mu_k^2)^2}(\frac{1}{2\cosh\frac{\sigma^k}{2}})^2 = 1\,, \\
g_{\mathcal{H}}(\boldsymbol{U}_{(k,l,2)}, \boldsymbol{U}_{(k,l,2)}) =& g_{\mathcal{H}}(\frac{1}{2\cosh\frac{\sigma^k}{2}\cosh\frac{\sigma^l}{2}}\frac{\boldsymbol{u}_{(k,l,2)}}{(\mu_k-\mu_l)}, \frac{1}{2\cosh\frac{\sigma^k}{2}\cosh\frac{\sigma^l}{2}}\frac{\boldsymbol{u}_{(k,l,2)}}{(\mu_k-\mu_l)}) \\
=& (\frac{1}{2\cosh\frac{\sigma^k}{2}\cosh\frac{\sigma^l}{2}(\mu_k-\mu_l)})^2\frac{4(\mu_k-\mu_l)^2}{(1-\mu_k^2)(1-\mu_l^2)} = 1\,.
\end{aligned}
$$

$\square$

2.3.3 Covariant derivatives

To determine all terms in SDE 2.2.8, we have the following Lemma for the expression of $\nabla_{\boldsymbol{V}}\boldsymbol{V}$ for all $\boldsymbol{L}$ and $\boldsymbol{U}$ vectors.

Lemma 8. *The covariant derivatives $\nabla_{\boldsymbol{U}}\boldsymbol{U}$ and $\nabla_{\boldsymbol{L}}\boldsymbol{L}$ are given by:*

$$
\begin{aligned}
\nabla_{\boldsymbol{L}_k}\boldsymbol{L}_k =& 0\,, && 1 \le k \le n\,, && (2.3.20) \\
\nabla_{\boldsymbol{U}_{(k)}}\boldsymbol{U}_{(k)} =& -\coth\sigma^k\boldsymbol{L}_k\,, && 1 \le k \le n\,, && (2.3.21) \\
\nabla_{\boldsymbol{U}_{(k,l,1)}}\boldsymbol{U}_{(k,l,1)} =& -\coth(\frac{\sigma^k}{2}+\frac{\sigma^l}{2})\frac{\boldsymbol{L}_k+\boldsymbol{L}_l}{2}\,, && 1 \le k < l \le n\,, && (2.3.22) \\
\nabla_{\boldsymbol{U}_{(k,l,2)}}\boldsymbol{U}_{(k,l,2)} =& -\coth(\frac{\sigma^k}{2}-\frac{\sigma^l}{2})\frac{\boldsymbol{L}_k-\boldsymbol{L}_l}{2}\,, && 1 \le k < l \le n\,. && (2.3.23)
\end{aligned}
$$

Proof. For a vector field $\boldsymbol{V} \in \mathfrak{X}(T\mathcal{H})$, We could compute $\nabla_{\boldsymbol{V}}\boldsymbol{V}$ by computing $\langle\nabla_{\boldsymbol{V}}\boldsymbol{V}, \boldsymbol{V}'\rangle$ using Koszul's formula [30, Theorem 3.6, Page 55], for $\boldsymbol{V}' = \sum_{k=1}^n a_k\boldsymbol{L}_k + \sum_{i=1}^{n^2} b_i\boldsymbol{U}_i$ with arbitrary constant coefficients a_k's and b_i's, where $\langle\cdot,\cdot\rangle = g_{\mathcal{H}}(\cdot,\cdot)$.

Koszul's formula is simplified in our case because all inner products between $\boldsymbol{U}$

and $\boldsymbol{L}$ vectors are constants. For example,

$$\langle \nabla_{\boldsymbol{U}_i}\boldsymbol{U}_i, \boldsymbol{L}_k\rangle = \boldsymbol{U}_i\langle \boldsymbol{U}_i, \boldsymbol{L}_k\rangle - \frac{1}{2}\boldsymbol{L}_k\langle \boldsymbol{U}_i, \boldsymbol{U}_i\rangle - \langle [\boldsymbol{U}_i, \boldsymbol{L}_k], \boldsymbol{U}_i\rangle = \langle [\boldsymbol{L}_k, \boldsymbol{U}_i], \boldsymbol{U}_i\rangle \,, \tag{2.3.24}$$

where the first two terms vanish because $\langle \boldsymbol{U}_i, \boldsymbol{U}_i\rangle = 1$ and $\langle \boldsymbol{U}_i, \boldsymbol{L}_k\rangle = 0$.

We give a proof of the expression of $\nabla_{\boldsymbol{L}_k}\boldsymbol{L}_k$ and $\nabla_{\boldsymbol{U}_{(k,l,2)}}\boldsymbol{U}_{(k,l,2)}$, then the computations are similar for others.

To prove $\nabla_{\boldsymbol{L}_k}\boldsymbol{L}_k = 0$, we need a fact discovered by Siegel in [83, Theorem 3, Page 3], implying that the translation of each σ^k is a geodesic. Also, notice that

$$\boldsymbol{L}_k(\sigma^l) = (2\cosh^2\frac{\sigma^l}{2})\boldsymbol{L}_k(\tanh\frac{\sigma^l}{2}) = (2\cosh^2\frac{\sigma^l}{2})\boldsymbol{L}_k(\mu_l) = \delta_{kl} \,, \tag{2.3.25}$$

which means $\boldsymbol{L}_k = \left(\frac{\partial}{\partial\sigma^k}\right)$, say $\boldsymbol{L}_k$ is the unit vector field tangent to the geodesic of σ^k translation. Therefore $\nabla_{\boldsymbol{L}_k}\boldsymbol{L}_k = 0$.

For $\boldsymbol{U}_{(k,l,2)}$, according to the formula above, we first need to distinguish those vector fields that are commutative with $\boldsymbol{U}_{(k,l,2)}$, or whose commutator with $\boldsymbol{U}_{(k,l,2)}$ are perpendicular to $\boldsymbol{U}_{(k,l,2)}$.

We first argue that $\langle \nabla_{\boldsymbol{U}_i}\boldsymbol{U}_i, \boldsymbol{U}_j\rangle = 0$ for all $\boldsymbol{U}_i, \boldsymbol{U}_j$ vectors. This is a result from (2.2.22): the commutator $[\boldsymbol{U}_i, \boldsymbol{U}_j]$ does not contain $\boldsymbol{U}_i$ or $\boldsymbol{U}_j$ term.

For $\langle \nabla_{\boldsymbol{U}_i}\boldsymbol{U}_i, \boldsymbol{L}_m\rangle$, $\boldsymbol{u}_{(k,l,2)}$ is by definition (2.2.29) commutative with $\boldsymbol{l}_m$ (thus $\boldsymbol{L}_m$), so

$$
\begin{aligned}
[\boldsymbol{L}_m, \boldsymbol{U}_{(k,l,2)}] =& [(\frac{\partial}{\partial \sigma^m}), \frac{1}{2\sinh(\frac{\sigma^k-\sigma^l}{2})}\boldsymbol{u}_{(k,l,2)}] \\
=& (\frac{\partial}{\partial \sigma^m}(\frac{1}{2\sinh(\frac{\sigma^k-\sigma^l}{2})}))\boldsymbol{u}_{(k,l,2)} \\
=& -\coth(\frac{\sigma^k-\sigma^l}{2})(\frac{\delta_{km}-\delta_{lm}}{2})\boldsymbol{U}_{(k,l,2)}\,.
\end{aligned}
\tag{2.3.26}
$$

Thus, $\nabla_{\boldsymbol{U}_{(k,l,2)}}\boldsymbol{U}_{(k,l,2)}$ is given by orthonormal decomposition below:

$$
\begin{aligned}
\nabla_{\boldsymbol{U}_{(k,l,2)}}\boldsymbol{U}_{(k,l,2)} =& \sum_{m=1}^{n} \langle \nabla_{\boldsymbol{U}_{(k,l,2)}}\boldsymbol{U}_{(k,l,2)}, \boldsymbol{L}_m\rangle \boldsymbol{L}_m \\
=& \sum_{m=1}^{n} \langle [\boldsymbol{L}_m, \boldsymbol{U}_{(k,l,2)}], \boldsymbol{U}_{(k,l,2)}\rangle \boldsymbol{L}_m \\
=& \sum_{m=1}^{n} -\coth(\frac{\sigma^k-\sigma^l}{2})(\frac{\delta_{km}-\delta_{lm}}{2})\boldsymbol{L}_m \\
=& -\coth(\frac{\sigma^k-\sigma^l}{2})(\frac{\boldsymbol{L}_k-\boldsymbol{L}_l}{2})\,.
\end{aligned}
\tag{2.3.27}
$$

□

2.3.4 Evolution of spectrum

Now it suffice to prove theorem 3, i.e. to compute the evolution of spectrum $\lambda(\boldsymbol{Z}_t) = (\tanh^2 \frac{\sigma^1}{2}, ..., \tanh^2 \frac{\sigma^n}{2})$.

Proof of theorem 1. Let $M_t = M(\boldsymbol{Z}_t)$. By definition (2.2.2), the evolution of M_t

satisfies the following SDE:

$$\begin{aligned}\mathrm{d}M_t =&\sum_{i=1}^{n^2}\Big(\boldsymbol{U}_i(M_t)\circ \mathrm{d}\tilde{W}_t^i-\frac{1}{2}\nabla_{\boldsymbol{U}_i}\boldsymbol{U}_i(M_t)\mathrm{d}t\Big)\\ &+\sqrt{\frac{2}{\beta}}\sum_{k=1}^{n}\Big(\boldsymbol{L}_k(M_t)\circ \mathrm{d}W_t^k-\frac{1}{2}\nabla_{\boldsymbol{L}_k}\boldsymbol{L}_k(M_t)\mathrm{d}t\Big)\\ =&\sum_{i=1}^{n^2}-\frac{1}{2}\nabla_{\boldsymbol{U}_i}\boldsymbol{U}_i(M_t)\mathrm{d}t+\sqrt{\frac{2}{\beta}}\sum_{k=1}^{n}\boldsymbol{L}_k(M_t)\circ \mathrm{d}W_t^k\,,\end{aligned}\tag{2.3.28}$$

where $\{\tilde{W}_t^i\}_{i=1}^{n^2}$ and $\{W_t^k\}_{k=1}^{n}$ are n^2+n independent Wiener processes.

Two facts were used in the computation above: $\boldsymbol{U}_i(\mathrm{d}M)=0$ for all $\boldsymbol{U}_i\in\{\boldsymbol{U}_i\}_{i=1}^{n^2}$ by definition (2.2.29), and $\nabla_{\boldsymbol{L}_k}\boldsymbol{L}_k=0$ for $k=1,...,n$ from Lemma 8.

Using the expression of $\nabla_{\boldsymbol{U}_i}\boldsymbol{U}_i$ in (2.3.20), and $\boldsymbol{L}_k(\mu_l)=\frac{1}{2\cosh^2\frac{\sigma^k}{2}}\delta_{kl}$ which is equivalent to $\boldsymbol{L}_k(M_t)=\frac{e_ke_k^T}{2\cosh^2\frac{\sigma^k}{2}}$, we can write down the SDE exactly as:

$$\begin{aligned}\mathrm{d}M_t =&\sum_{k=1}^{n}\mathrm{d}\big(\tanh(\frac{\sigma^k}{2})\big)e_ke_k^T\\ =&\sum_{k=1}^{n}\frac{1}{2}\coth(\sigma_t^k)\frac{e_ke_k^T}{2\cosh^2\frac{\sigma^k}{2}}\mathrm{d}t+\sum_{1\le k<l\le n}\frac{1}{2}\coth(\frac{\sigma_t^k+\sigma_t^l}{2})\frac{\frac{e_ke_k^T}{2\cosh^2\frac{\sigma^k}{2}}+\frac{e_le_l^T}{2\cosh^2\frac{\sigma^l}{2}}}{2}\mathrm{d}t\\ &+\sum_{1\le k<l\le n}\frac{1}{2}\coth(\frac{\sigma_t^k-\sigma_t^l}{2})\frac{\frac{e_ke_k^T}{2\cosh^2\frac{\sigma^k}{2}}-\frac{e_le_l^T}{2\cosh^2\frac{\sigma^l}{2}}}{2}\mathrm{d}t+\sqrt{\frac{2}{\beta}}\sum_{k=1}^{n}\frac{e_ke_k^T}{2\cosh^2\frac{\sigma^k}{2}}\circ \mathrm{d}W_t^k\\ =&\sum_{k=1}^{n}\Big(\frac{1}{2\cosh^2(\frac{\sigma^k}{2})}\times\frac{1}{2}\big(\coth(\frac{\sigma_t^k}{2})+\sum_{l\neq k}\frac{\sinh(\sigma_t^k)}{\cosh(\sigma_t^k)-\cosh(\sigma_t^l)}\big)\mathrm{d}t\\ &+\sqrt{\frac{2}{\beta}}\frac{1}{2\cosh^2(\frac{\sigma^k}{2})}\circ \mathrm{d}W_t^k\Big)e_ke_k^T\end{aligned}\tag{2.3.29}$$

As $\{e_ke_k^T\}_{k=1}^{n}$ are linearly independent, the coefficients on the two sides coincide, and we get the evolution of $\tanh(\frac{\sigma_t^k}{2})$. Using chain rule we find the evolution of σ_t^k:

$$\begin{aligned}
\mathrm{d}\sigma_t^k =&\frac{1}{2}\big(\coth(\frac{\sigma_t^k}{2})+\sum_{l\neq k}\frac{\sinh(\sigma_t^k)}{\cosh(\sigma_t^k)-\cosh(\sigma_t^l)}\big)\mathrm{d}t+\sqrt{\frac{2}{\beta}}\times 1\circ \mathrm{d}W_t^k\\
=&\frac{1}{2}\big(\coth(\frac{\sigma_t^k}{2})+\sum_{l\neq k}\frac{\sinh(\sigma_t^k)}{\cosh(\sigma_t^k)-\cosh(\sigma_t^l)}\big)\mathrm{d}t+\sqrt{\frac{2}{\beta}}\mathrm{d}W_t^k
\end{aligned}$$

which is (2.1.7). $\square$

2.4 Proof of Theorem 4

In this section, we evaluate the Boltzmann entropy $S : \mathcal{Q} \to \mathbb{R}$ defined in equation (2.1.19) and prove Theorem 4.

2.4.1 Orbit volume

For each $\lambda \in \mathcal{Q}$, the orbit $\mathcal{O}_\lambda$ is naturally a Riemannian submanifold of $\mathcal{H}$ equipped with the induced metric $g_{\mathcal{O}_\lambda}$, defined by

$$g_{\mathcal{O}_\lambda}(\boldsymbol{v}, \boldsymbol{v}') = g_{\mathcal{H}}(\boldsymbol{v}, \boldsymbol{v}') \,, \tag{2.4.1}$$

for $\boldsymbol{v}, \boldsymbol{v}' \in T\mathcal{O}_\lambda$.

We first introduce a collection of n^2 1-forms to be the basis of $\mathfrak{X}(T^*\mathcal{O}_\lambda)$. Denoted $\{\omega^i\}_{i=1}^{n^2} \subset \mathfrak{X}(T^*\mathcal{O}_\lambda)$ that are dual to the orthogonal basis $\{\boldsymbol{u}^i\}_{i=1}^{n^2}$ defined in (2.2.25). That is,

$$\boldsymbol{u}^i(\omega^j) = \delta_{ij} \,, \qquad\qquad i, j = 1, ..., n^2 \,, \tag{2.4.2}$$

using (2.3.19) the metric $g_{\mathcal{O}_\lambda}$ has the form

$$\begin{aligned}
g_{\mathcal{O}_\lambda} =& \sum_{k=1}^{n} \frac{4(2\mu_k)^2}{(1-\mu_k^2)^2} \omega^{(k)} \otimes \omega^{(k)} \\
&+ \sum_{1\le k<l\le n} \frac{4(\mu_k+\mu_l)^2}{(1-\mu_k^2)(1-\mu_l^2)} \omega^{(k,l,1)} \otimes \omega^{(k,l,1)} \\
&+ \sum_{1\le k<l\le n} \frac{4(\mu_k-\mu_l)^2}{(1-\mu_k^2)(1-\mu_l^2)} \omega^{(k,l,2)} \otimes \omega^{(k,l,2)} \\
=& 4\Big(\sum_{k=1}^{n} \sinh^2 \sigma^k \omega^{(k)} \otimes \omega^{(k)} + \sum_{1\le k<l\le n} \sinh^2(\frac{\sigma^k+\sigma^l}{2}) \omega^{(k,l,1)} \otimes \omega^{(k,l,1)} \\
&+ \sum_{1\le k<l\le n} \sinh^2(\frac{\sigma^k-\sigma^l}{2}) \omega^{(k,l,2)} \otimes \omega^{(k,l,2)} \Big).
\end{aligned} \tag{2.4.3}$$

Let $M_{\lambda,\omega}$ be the matrix of $g_{\mathcal{O}_\lambda}$ under basis $\{\omega^i\}_{i=1}^{n^2}$, then the volume form is $\mathrm{d}V = \sqrt{\det(M_{\lambda,\omega})} \bigwedge_{i=1}^{n^2} \omega^i$ and $\mathrm{vol}(\mathcal{O}_\lambda) = \int_{\mathcal{O}_\lambda} \mathrm{d}V$ gives the volume of orbit $\mathcal{O}_\lambda$.

Notice that $\det(M_{\lambda,\omega})$ only depends on λ, and is a constant on $\mathcal{O}_\lambda$, we can extract it out and get

$$\mathrm{vol}(\mathcal{O}_\lambda) = 2^{n^2} \prod_{k=1}^{n} \sinh \sigma^k \prod_{1\le k<l\le n} \sinh(\frac{\sigma^k+\sigma^l}{2}) \sinh(\frac{\sigma^k-\sigma^l}{2}) \int_{\mathcal{O}_\lambda} \bigwedge_{i=1}^{n^2} \omega^i \tag{2.4.4}$$

and we will argue that the integral $\int_{\mathcal{O}_\lambda} \bigwedge_{i=1}^{n^2} \omega^i$ is a constant that does not depend on λ, so we have the following

Theorem 10. *For $\lambda \in \mathcal{Q}$, the volume of orbit $\mathcal{O}_\lambda$ is given by*

$$\mathrm{vol}(\mathcal{O}_\lambda) = C_n \prod_{k=1}^{n} \sinh \sigma^k \prod_{1\le k<l\le n} \sinh(\frac{\sigma^k+\sigma^l}{2}) \sinh(\frac{\sigma^k-\sigma^l}{2}) \tag{2.4.5}$$

where C_n is a constant that only depends on n.

Proof. We only need to show that the integral $\int_{\mathcal{O}_\lambda} \bigwedge_{i=1}^{n^2} \omega^i = \int_{\mathcal{O}'_\lambda} \bigwedge_{i=1}^{n^2} \omega^i$ for $\lambda, \lambda' \in \mathcal{Q}$.

To do this we consider a bijection

$$\phi : \mathcal{O}_\lambda \to \mathcal{O}_{\lambda'}\,, \text{ for } Z = \varphi^{-1}(QMQ^T) \in \mathcal{O}_\lambda\,,\, \phi(Z) = \varphi^{-1}(QM'Q^T)\,.$$

i.e. ϕ is the map that preserves the $\mathrm{U}(n)$ component of R coordinate.

Since ϕ is a diffeomorphism between $\mathcal{O}_\lambda$ and $\mathcal{O}_{\lambda'}$, using $\phi^*(\omega^i \circ \phi) \in \mathfrak{X}(T^*\mathcal{O}_\lambda)$, the pullback of $\omega^i \circ \phi : \mathcal{O}_\lambda \to T^*\mathcal{O}_{\lambda'}$, we can use change of variable formula to get

$$\begin{aligned} \int_{\mathcal{O}_{\lambda'}} \bigwedge_{i=1}^{n^2} \omega^i &= \int_{\phi(\mathcal{O}_\lambda)} \bigwedge_{i=1}^{n^2} \omega^i \\ &= \int_{\mathcal{O}_\lambda} \bigwedge_{i=1}^{n^2} \phi^*(\omega^i \circ \phi)\,, \end{aligned} \tag{2.4.6}$$

Now we will show that $\phi^*(\omega^i \circ \phi) = \omega^i$. This is the direct result of $\phi_*(\boldsymbol{u}_i) = \boldsymbol{u}_i \circ \phi$, to prove which consider $Z' = \phi(Z)$ for $Z \in \mathcal{O}_\lambda$ and an arbitrary function $f \in C^\infty(\mathcal{O}_{\lambda'})$, then we need to prove that $\phi_*(\boldsymbol{u}_i(Z))f = \boldsymbol{u}_i(Z')f$.

Let the matrix-valued derivative $\frac{\partial f}{\partial Q}$ defined by $\mathrm{d}f = \mathrm{Tr}(\frac{\partial f}{\partial Q}\mathrm{d}Q)$, we have

$$\begin{aligned} \phi_*(\boldsymbol{u}_i(Z))f =& \boldsymbol{u}_i(f \circ \phi(Z)) = \mathrm{Tr}\Big((\frac{\partial f}{\partial Q})_{Z'}(\boldsymbol{u}_i(\mathrm{d}Q))_Z\Big)\,, \\ \boldsymbol{u}_i(Z')f =& \mathrm{Tr}\Big((\frac{\partial f}{\partial Q})_{Z'}(\boldsymbol{u}_i(\mathrm{d}Q))_{Z'}\Big)\,, \end{aligned}$$

where we used the fact that Z and Z' has the same Q component. As $(\boldsymbol{u}_i(\mathrm{d}Q))_{Z'} = (\boldsymbol{u}_i(\mathrm{d}Q))_Z$ from (2.2.29), we proved that $\phi_*(\boldsymbol{u}_i(Z))f = \boldsymbol{u}_i(Z')f$.

Then by

$$\boldsymbol{u}_j(\phi^*(\omega^i \circ \phi)) = \phi_*(\boldsymbol{u}_j)(\omega^i \circ \phi) = (\boldsymbol{u}_i \circ \phi)(\omega^i \circ \phi) = \delta^i_j\,, \tag{2.4.7}$$

we proved that $\phi^*(\omega^i \circ \phi) = \omega^i$.

So the integral is a constant:

$$\int_{\mathcal{O}_{\lambda'}} \bigwedge_{i=1}^{n^2} \omega^i = \int_{\mathcal{O}_\lambda} \bigwedge_{i=1}^{n^2} \phi^*(\omega^i \circ \phi) = \int_{\mathcal{O}_\lambda} \bigwedge_{i=1}^{n^2} \omega^i \,. \tag{2.4.8}$$

$\square$

Take logarithm and using the formula $\sinh(\frac{a+b}{2})\sinh(\frac{a-b}{2}) = \frac{\cosh(a)-\cosh(b)}{2}$, we get the Boltzmann entropy

$$S(\sigma) = \log \mathrm{vol}(\mathcal{O}_\lambda) = \sum_{k=1}^{n} \log\sinh\sigma^k + \sum_{1\le k<l\le n} \log|\cosh\sigma^k - \cosh\sigma^l| + c_n \,. \tag{2.4.9}$$

2.4.2 Metric and gradient flow

We have already computed the Boltzmann entropy function $S(\sigma)$, and to find a gradient flow we will also need a metric g_Λ in the spectrum space $\mathcal{Q}$. The natural selection is to make the projection $\pi : \mathcal{H} \to \mathcal{Q}$ a Riemannian submersion.

Now as we already have an orthonormal basis of $T\mathcal{H}$ and it is clear that

$$\mathrm{Ker}(\mathrm{d}\pi) = \mathrm{Span}\Big(\{\boldsymbol{U}_i\}_{i=1}^{n^2}\Big)\,, \qquad \mathrm{Ker}(\mathrm{d}\pi)^\perp = \mathrm{Span}\Big(\{\boldsymbol{L}_k\}_{k=1}^{n}\Big)\,. \tag{2.4.10}$$

By definition of a Riemannian submersion, $\{\mathrm{d}\pi(\boldsymbol{L}_k)\}_{k=1}^n$ will be an orthonormal basis in $T\mathcal{Q}$. Using (2.3.25) we have

$$g_{\mathcal{Q}} = \sum_{k=1}^{n} \mathrm{d}\sigma^k \mathrm{d}\sigma^k\,, \qquad (g_{\mathcal{Q}})^{-1} = \sum_{k=1}^{n} \boldsymbol{L}_k \otimes \boldsymbol{L}_k \,. \tag{2.4.11}$$

Therefore, the gradient of $S(\sigma)$ is just the usual differential $\mathrm{grad}\, S(\mathrm{d}\sigma^k) = \frac{\partial S}{\partial \sigma^k}$. Comparing SDE (2.1.7) and the entropy (2.1.20) we conclude that the evolution of $(\sigma^1, ..., \sigma^n)$ is just the Langevin equation maximizing Boltzmann entropy S.

2.5 Well-posedness and Non-colliding property

In this section, we prove the existence and uniqueness of strong solution of SDE (2.1.7), when $\beta \geq 2$ and the initial condition satisfies $\lambda_0 \in \mathcal{Q}$, i.e. $\sigma_0 = (\sigma_0^i)_{i=1}^n$ are all distinct and non-zero. The essential observation is that Boltzmann entropy $S(\sigma_t)$ is a local submartingale and with probability 1 it is bounded below, which avoids collision of eigenvalues. This strategy is commonly used for the well-posedness of Dyson Brownian motion in random matrix theory, and our proof adapts arguments used in [2, 8].

Theorem 11. *When $\beta \geq 2$, with initial condition $\sigma_0 = (\sigma_0^i)_{i=1}^n$ such that $S(\sigma_0) > -\infty$, there exists a unique strong solution to the SDE* (2.1.7)*, and $S(\sigma_t) > -\infty$ for all $t > 0$ almost surely.*

Our overall strategy is to consider an auxiliary SDE with a cut-off of the coefficients in (2.1.7) such that well-posedness is guaranteed, and properly define some stopping times so that the solution to the auxiliary SDE also solves SDE (2.1.7) before they stopped. We then define the solution of SDE (2.1.7) to be equal to the solution of auxiliary SDE before stopping, then we let the stopping times goes to infinity to obtain a solution almost surely defined for all $t > 0$.

The auxiliary SDE is as follows: define $N(\sigma) = \log(\sum_{i=1}^n \cosh \sigma^i)$, for $k > -S(\sigma_0), K > N(\sigma_0)$, let $\sigma_t^{(k,K)}$ solves the SDE as follows:

$$\mathrm{d}(\sigma_t^{(k,K)})^i = \eta(\sigma_t^{(k,K)}, k, K)\Big(\frac{1}{2}(\coth(\sigma_t^{(k,K)})^i + \sum_{j\neq i} \frac{\sinh(\sigma_t^{(k,K)})^i}{\cosh(\sigma_t^{(k,K)})^i - \cosh(\sigma_t^{(k,K)})^j})\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}W_t^i\Big), \tag{2.5.1}$$

where $\eta(\sigma, k, K) = h(\frac{-S(\sigma)}{k})h(\frac{N(\sigma)}{K})$ and h is a smooth function with

$$h(x) = \begin{cases} 1 & x \leq 1 \\ \\ 0 & x \geq 2 \end{cases} . \tag{2.5.2}$$

Thus, when $S(\sigma) > -k$, $N(\sigma) < K$, the SDE (2.5.1) and the SDE (2.1.7) coincides.

Now define two stopping times

$$\tau_{k,K} := \inf\{t > 0 | S(\sigma_t^{(k,K)}) \leq -k\}\,, \tag{2.5.3}$$

$$T_{k,K} := \inf\{t > 0 | N(\sigma_t^{(k,K)}) \geq K\}\,. \tag{2.5.4}$$

Notice that when $S(\sigma) > -k$ and $N(\sigma) < K$, $\sigma^i, S(\sigma), \frac{\partial S}{\partial \sigma^i}$ are all bounded, where $\frac{\partial S}{\partial \sigma^i} = \mathrm{e}^{-S}\frac{\partial(\mathrm{e}^S)}{\partial \sigma^i}$ is bounded because e^S is smooth in $\{\sigma \in \mathbb{R}^n | N(\sigma) \leq K\}$. We now have the following two Lemmas for $T_{k,K}$ and $\tau_{k,K}$.

Lemma 9. *For $t > 0, k > -S(\sigma_0)$, $\lim_{K\to\infty} \mathbb{P}[T_{k,K} < t \wedge \tau_{k,K}] = 0$ uniformly in k.*

Lemma 10. *For $K > N(\sigma_0), t > 0$, $\lim_{k\to\infty} \mathbb{P}[\tau_{k,K} < T_{k,K} \wedge t] = 0$.*

Corollary 2. *Given $t > 0$, almost surely there exists $k_* > -S(\sigma_0), K_* > N(\sigma_0)$ such that $t < \tau_{k_*,K_*} \wedge T_{k_*,K_*}$.*

Proof. For each $n \in \mathbb{N}$, there exists K_n such that

$$\mathbb{P}[T_{k,K_n} < t \wedge \tau_{k,K_n}] < 2^{-n}\,, \tag{2.5.5}$$

for all $k > -S(\sigma_0)$ by Lemma 9. There also exists k_n such that

$$\mathbb{P}[\tau_{k_n,K_n} < t \wedge T_{k_n,K_n}] < 2^{-n}\,, \tag{2.5.6}$$

by Lemma 10.

Since $\mathbb{P}[t > T_{k_n,K_n} \wedge \tau_{k_n,K_n}] \leq \mathbb{P}[\tau_{k_n,K_n} < t \wedge T_{k_n,K_n}] + \mathbb{P}[T_{k_n,K_n} < t \wedge \tau_{k_n,K_n}]$

$$\sum_{n=1}^{\infty} \mathbb{P}[t > T_{k_n,K_n} \wedge \tau_{k_n,K_n}] < \infty \tag{2.5.7}$$

by the Borel-Cantelli Lemma, with probability 1, $\{t > T_{k_n,K_n} \wedge \tau_{k_n,K_n}\}_{n=1}^{\infty}$ happens finitely often. Hence, there exist an integer m such that $t \leq T_{k_m,K_m} \wedge \tau_{k_m,K_m}$. $\square$

Corollary 2 leads to proof of Theorem 11 directly:

Proof of Theorem 11. For all $k > -S(\sigma_0), K > N(\sigma_0)$, SDE (2.5.1) has a unique strong solution $\sigma_t^{(k,K)}$ because coefficients of drift and diffusion are all Lipschitz.

For each $t > 0$, almost surely there exists k_*, K_* such that $t < \tau_{k_*,K_*} \wedge T_{k_*,K_*}$, we define $\sigma_t = \sigma_t^{(k_*,K_*)}$. By uniqueness of solution of SDE (2.5.1), such definition is well-defined and does not depend on the choice of k_*, K_*, because $\sigma_t^{(k,K)} = \sigma_t^{(k',K')}$ when $t \leq \tau_{k,K} \wedge T_{k,K}$ and $t \leq \tau_{k',K'} \wedge T_{k',K'}$.

It is clear that σ_t is a solution to SDE (2.1.7), and uniqueness follows the uniqueness of $\sigma_t^{(k,K)}$ of SDE (2.5.1).

The finiteness of $S(\sigma_t)$ comes from that $t < \tau_{k_*,K_*}$ which means $S(\sigma_t) > -k_*$. $\square$

The proof of Lemma 9 and Lemma 10 are as follows:

Proof of Lemma 9. Consider $N_t^{k,K} = N(\sigma_t^{(k,K)})$, using Itô's formula, when $t \leq \tau_{k,K} \wedge T_{k,K}$, we have

$$\begin{aligned} N_t^{k,K} =& N(\sigma_0) + \sqrt{\frac{2}{\beta}} \int_0^t \sum_{i=1}^n \frac{\sinh(\sigma_s^{(k,K)})^i}{\sum_{l=1}^n \cosh(\sigma_s^{k,K})^l} \mathrm{d}W_s^i \\ &+ \int_0^t \left(\frac{n}{2} + \frac{1}{\beta} - \frac{1}{\beta} \frac{\sum_{i=1}^n \sinh^2(\sigma_s^{(k,K)})^i}{(\sum_{i=1}^n \cosh(\sigma_s^{(k,K)})^i)^2}\right) \mathrm{d}s\,. \end{aligned} \tag{2.5.8}$$

As $\sigma_s^{(k,K)}$ is bounded for $0 \le s \le t \wedge \tau_{k,K} \wedge T_{k,K}$, each term on the R.H.S are integrable. Take expectation of $N_{t\wedge\tau_{k,K}\wedge T_{k,K}}$:

$$\begin{aligned} K\mathbb{P}[T_{k,K} < t \wedge \tau_{k,K}] \le& \mathbb{E}[N_{t\wedge\tau_{k,K}\wedge T_{k,K}}] \\ \le& N(\sigma_0) + (\frac{n}{2} + \frac{1}{\beta})\mathbb{E}[\int_0^{t\wedge\tau_{k,K}\wedge T_{k,K}} \mathrm{d}s] \\ =& N(\sigma_0) + (\frac{n}{2} + \frac{1}{\beta})\mathbb{E}[\int_0^{t\wedge\tau_{k,K}\wedge T_{k,K}} \mathrm{d}s] \\ \le& N(\sigma_0) + (\frac{n}{2} + \frac{1}{\beta})t\,, \end{aligned} \tag{2.5.9}$$

so $\lim_{K\to\infty} \mathbb{P}[T_{k,K} < t \wedge \tau_{k,K}] = 0$ uniformly in k.

□

Proof of Lemma 10. We show that when $\beta \ge 2$, $S(\sigma_{t\wedge\tau_{k,K}\wedge T_{k,K}}^{(k,K)})$ is a submartingale. We start with the following observation:

$$\Delta S + |\nabla S|^2 = c_n\,, \tag{2.5.10}$$

where $\Delta = \sum_{i=1}^n \frac{\partial^2}{\partial\sigma^i\partial\sigma^i}$ is the Laplacian in Euclidean sense, $|\nabla S|^2 = \sum_{i=1}^n (\frac{\partial S}{\partial\sigma^i})^2$, and $c_n = \frac{n(n+1)(2n+1)}{6} > 0$.

When $t \le \tau_{k,K} \wedge T_{k,K}$, using Itô's formula and SDE (2.1.7), we have

$$\begin{aligned} S(\sigma_t^{(k,K)}) =& S(\sigma_0) + \int_0^t (\frac{1}{2}|\nabla S|^2 + \frac{1}{\beta}\Delta S)\mathrm{d}s + \int_0^t \sum_{i=1}^n \sqrt{\frac{2}{\beta}}\frac{\partial S}{\partial\sigma^i}\mathrm{d}W_s^i \\ =& S(\sigma_0) + \int_0^t (\frac{c_n}{\beta} + (\frac{1}{2} - \frac{1}{\beta})|\nabla S|^2)\mathrm{d}s + \int_0^t \sum_{i=1}^n \sqrt{\frac{2}{\beta}}\frac{\partial S}{\partial\sigma^i}\mathrm{d}W_s^i \end{aligned} \tag{2.5.11}$$

When $0 \le s \le t \wedge \tau_{k,K} \wedge T_{k,K}$, every integral above are integrable. Using the condition $\beta \geq 2$, we have

$$S(\sigma_0) < \mathbb{E}[S_{t\wedge\tau_{k,K}\wedge T_{k,K}}] \tag{2.5.12}$$

Also, $S_{t\wedge\tau_k\wedge T_K}$ is bounded above by a constant C_K depending on K, we have

$$S_0 < \mathbb{E}[S_{t\wedge\tau_{k,K}\wedge T_{k,K}}] < -k\mathbb{P}[\tau_{k,K} < t\wedge T_{k,K}] + C_K(1 - \mathbb{P}[\tau_{k,K} < t\wedge T_{k,K}])\,, \tag{2.5.13}$$

which leads to $\lim_{k\to\infty}\mathbb{P}[\tau_{k,K} < t\wedge T_{k,K}] = 0$ for given $t > 0, K > N(\sigma_0)$.

$\square$

Chapter 3

Gibbs sampling problem on PSD matrices

3.0 Abstract

We construct Riemannian Langevin equations (RLE) on the manifold of fixed rank positive semi-definite matrices, for a suitable smooth energy function under two classical Riemannian metrics. One metric is associated with the embedded geometry in Euclidean space, and the other one is Bures-Wasserstein metric. Based on the interplay of mean curvature and intrinsic Brownian motion in Riemannian immersion and submersion, we use orthonormal global smooth frames for parallelizable manifolds to give a simpler construction of Brownian motion, which is different from the classical Eells-Elworth-Malliavin approach yet more useful for obtaining Itô form of Brownian motion toward constructing efficient Riemannian Langevin Monte Carlo schemes.

3.1 Introduction

3.1.1 Problem statement

We consider the stochastic process X_t on the manifold of positive semi-definite (PSD) matrices of fixed rank $\mathcal{S}_+^{n,p}$ defined by

$$\mathcal{S}_+^{n,p} = \left\{ X \in \mathbb{R}^{n\times n} \middle| X = X^T, X \succ 0, \mathrm{Rank}(X) = p \right\}, \tag{3.1.1}$$

solving the following Riemannian Langevin equation (RLE)

$$\mathrm{d}X_t = -\mathrm{grad}\,\mathcal{E}\mathrm{d}t + \mathrm{d}\boldsymbol{B}_t^{g,\beta}, \tag{3.1.2}$$

For a given energy function $\mathcal{E} \in C^2(\mathcal{S}_+^{n,p})$ and a metric g equipped to $\mathcal{S}_+^{n,p}$, where $\boldsymbol{B}_+^{g,\beta}$ stands for the intrinsic Brownian motion on the Riemannian manifold $(\mathcal{S}_+^{n,p}, g)$, and grad is Riemannian gradient, $\beta > 0$ is inverse temperature.

RLE (3.1.2) connects to both Gibbs sampling problem and optimization problem on $\mathcal{S}_+^{n,p}$. The equilibrium distribution of RLE (3.1.2) is Gibbs measure

$$\rho_* = \frac{1}{Z_\beta}\mathrm{e}^{-\beta\mathcal{E}}, \qquad Z_\beta = \int_{\mathcal{S}_+^{n,p}} \mathrm{e}^{-\beta\mathcal{E}}\mathrm{d}V, \tag{3.1.3}$$

where $\mathrm{d}V = \sqrt{\det(g)}\mathrm{d}x$ is the volume form and we require $Z_\beta < \infty$. As the metric appears in the volume form, Gibbs measures associated to the same energy function $\mathcal{E}$ on $\mathcal{S}_+^{n,p}$ equipped with different metrics are different. In the zero temperature limit $\beta \to \infty$, RLE (3.1.2) reduces to gradient descent and ρ_* concentrates to the global minima of $\mathcal{E}$.

The main contributions of this paper are two simple and explicit forms of RLE

14 15 on $\mathcal{S}_+^{n,p}$ under g_E induced metric from Euclidean metric, and g_{BW} the Bures-Wasserstein metric, companioned with convergence analysis based on Bakry-Émery criteria, that provide new methods doing Gibbs sampling and optimization problem on the manifold $\mathcal{S}_+^{n,p}$. Also, both RLEs can be extended to general Riemannian manifolds with certain conditions. The most subtle part in RLE is the construction of intrinsic Brownian motion on Riemannian manifolds which has been well studied in literature, e.g. [43, 47, 84, 10]. Different from the usual Eells-Elworth-Malliavin construction through horizontal lift to frame bundle, in this paper we assumed stronger conditions and provide a construction which is simpler and more practicable for numerical applications, see 12. Further more, for the specific case $\mathcal{S}_+^{n,p}$, we are taking advantage of its structure (as an embedded manifold or a quotient manifold) and providing even simpler SDEs for Brownian motions on it 12 13. The key phenomenon here is the interplay between Brownian motion and (negative) mean curvature flow, as well as gradient ascent of log volume of group orbit which is called a Boltzmann entropy in this paper. These connections appear in literature [50, 84, 78] while less well-known.

In a separate paper [94], our Riemannian Langevin schemes are reported. The foundation of the Riemannian Langevin schemes in [94] is the Riemannian Langevin equation, which is the focus of this paper.

The technical tools used in this paper includes stochastic analysis on manifolds (see [43]) and Riemannian geometry (see [30]). Corresponding definition and notations are also introduced in section 3.2, as well as main theorems. In section 3.4 we discussed the convergence based on classical Bakry-Émery criteria for two metrics respectively, with concrete computation for curvature tensor of $(\mathcal{S}_+^{n,p}, g_E)$.

3.1.2 Two special Riemannian metrics of $S_+^{n,p}$

$\mathcal{S}_+^{n,p}$ as a constrained system is naturally a submanifold of $\mathrm{Sym}(n) = \{X \in \mathbb{R}^{n\times n} | X = X^T\}$, the collection of all $n \times n$ positive semi-definite matrices. $\mathrm{Sym}(n)$ can be equipped with Euclidean metric $g(A, B) = \mathrm{Tr}(AB^T)$, where A, B are two vectors in $T\mathrm{Sym}(n)$ and Tr is the trace of matrices. The induced metric g_E, defined by the restriction of g on $T\mathcal{S}_+^{n,p}$, gives the natural metric structure on $\mathcal{S}_+^{n,p}$ as a submanifold [91]. i.e.

$$g_{E,X}(A, B) = g(A, B) = \mathrm{Tr}(AB^T)\,, \qquad A, B \in T_X\mathcal{S}_+^{n,p}\,. \tag{3.1.4}$$

This metric g_E is also called embedded metric in this paper.

On the other hand, $\mathcal{S}_+^{n,p}$ is naturally (more precisely, is homeomorphic to) a quotient manifold of $\mathbb{R}_*^{n\times p}$, where $\mathbb{R}_*^{n\times p}$ is the collection of full rank $n \times p$ matrices. This can be understood through the decomposition $X = YY^T$ for $X \in \mathcal{S}_+^{n,p}$, $Y \in \mathbb{R}_*^{n\times p}$. By the invariance of X under the action $Y \to YQ$, $Q \in \mathcal{O}_p$, the orthogonal group, we can see $\mathcal{S}_+^{n,p} \approx \mathbb{R}_*^{n\times p}/\mathcal{O}_p$.

Equip $\mathbb{R}_*^{n\times p}$ with Euclidean metric $g(U, V) = \mathrm{Tr}(UV^T)$, g induces a metric on the quotient $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$ through Riemannian submersion for the natural projection. This metric pulling back to $\mathcal{S}_+^{n,p}$, denoted by g_{BW}, is known as Bures-Wasserstein metric [70, 69]. The metric g_{BW} is also called quotient metric in this paper.

3.1.3 Riemannian Langevin equation

Riemannian Langevin equation (RLE) on a Riemannian manifold $(\mathcal{M}, g)$ is of the following form:

$$\mathrm{d}\boldsymbol{X}_t = -\mathrm{grad}\,\mathcal{E}\mathrm{d}t + \mathrm{d}\boldsymbol{B}_t^{g,\beta}\,, \tag{3.1.5}$$

where $\boldsymbol{X}_t$ is a $\mathcal{M}$-valued process, grad is Riemannian gradient, $\boldsymbol{B}_t^{g,\beta}$ is the intrinsic Brownian motion on the manifold $(\mathcal{M}, g)$ with temperature $\frac{1}{\beta}$. SDEs on manifolds are defined in the weak sense [43], i.e. (3.1.2) defines a process with infinitesimal generator $\mathcal{L}f = -\mathrm{grad}\,\mathcal{E}(f) + \frac{1}{\beta}\Delta_g f$, where Δ_g is the Laplace-Beltrami operator under metric g.

RLE (3.1.2) is a natural extension of the following Langevin equation in Euclidean space

$$\mathrm{d}x_t = -\nabla\mathcal{E}\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}W_t\,, \tag{3.1.6}$$

where W_t is a $\mathbb{R}^n$ valued Brownian motion. To extend this equation to Riemannian manifolds, Euclidean coordinates x is replaced by local coordinates on manifolds, so the whole SDE should be written in a coordinate free form. The gradient $\nabla\mathcal{E}$ is replaced by Riemannian gradient, and Brownian motion W_t should also be replaced by intrinsic Brownian motion on Riemannian manifolds. For this sake, SDE on manifolds are written in the coordinate-free Stratonovich form. On the other hand, to be a martingale on manifolds, $\boldsymbol{B}_t^{g,\beta}$ should satisfy a Itô form SDE, which makes Itô-Stratonovich correction naturally appear in SDE of $\boldsymbol{B}_t^{g,\beta}$, e.g.12. Such correction term possesses a meaning of mean curvature in the situation where $\mathcal{M}$ is a embedded manifold in its ambient space 12, and gives a drift of entropy gradient in the situation where $\mathcal{M}$ is a quotient manifold 13. This phenomenon has been studied in depth by the authors (TY and GM) and their co-authors in literature [71, 72, 48].

The Fokker-Planck equation describing the density ρ_t of $\boldsymbol{X}_t$ also extends to Riemannian case naturally. Density ρ of $\boldsymbol{X}$ on a Riemannian manifolds are defined in the following sense, for any $f \in C^\infty(\mathcal{M})$,

$$\mathbb{E}[f(\boldsymbol{X})] = \int_{\mathcal{M}} \rho(X)f(X)\mathrm{d}V = \int_{\Omega} \rho(x)f(x)\sqrt{\det(g_{ij})}\mathrm{d}x\,, \tag{3.1.7}$$

where $x : \mathcal{M} \to \Omega$ is a coordinate chart and g_{ij} is the matrix of g, $\mathrm{d}V = \sqrt{\det(g_{ij})}$ is the volume form. Because of the dependence of $\mathrm{d}V$ on the metric, densities on the same manifold with different metric may give different value of $\mathbb{E}[f]$.

Fokker-Planck equation for ρ_t is

$$\partial_t \rho_t = \mathrm{div}(\frac{1}{\beta}\mathrm{grad}\rho_t + \rho_t\,\mathrm{grad}\mathcal{E}) = \frac{1}{\sqrt{g}}\partial_i(\sqrt{g}g^{ij}(\frac{1}{\beta}\partial_j\rho_t + \rho_t\partial_j\mathcal{E}))\,, \tag{3.1.8}$$

and will be discussed in 3.4. Fokker-Planck type equation has been studied extensively in literature, but the considerations on Riemannian manifolds are limited. e.g. the equation 3.1.8 lies in the situation where coefficients are varying and the elliptic condition $g^{ij}\xi_i\xi_j > \kappa|\xi|^2$ for some constant $\kappa > 0$ may not be satisfied, which is beyond a common assumption in existing literature.

Though the Brownian motion on manifolds can be also constructed by the classical Eells-Elworthy-Malliavin's approach ,e.g., [32, 66, 84, 43, 10], our approach gives an explicit expression, which can be easily implemented for numerical schemes.

3.1.4 Connection to existing works and contributions

PSD fixed rank matrices are related to matrices in many problems such as distance matrices [88] and covariance matrices in statistics, and have been used in applications including kernels in machine learning [73], semidefinite optimization [19], etc. Classical Riemannian optimization algorithms over $S_+^{n,p}$ under different metrics have been well studied, e.g., see [91, 53, 70, 96] and references therein.

For the two metrics on $\mathcal{S}_+^{n,p}$, Embedded geometry g_E was studied in [91], BW metric was mentioned in [53, 70]. There are also other metrics considered on $\mathcal{S}_+^{n,p}$, e.g. see [70, 96].

Studies on Riemannian Langevin algorithms arises in recent years [22, 23, 35, 62], and all of them use exponential map to implement Brownian motion. In a separate paper [94] concerning numerical implementation of RLE introduced in this paper, we simply used Euler-Murayama discretization to do MCMC. There are also other Langevin-based Markov Chain Monte Carlo schemes as long as Hamiltonian Monte Carlo schemes on manifolds studied in literature [38, 20, 74, 95, 92]. Langevin dynamics with simulated annealing is also an established approach [65].

Standard Fokker-Planck equation is a well-studied object, see e.g.[15, 33, 14] on the well-posedness problem. On the convergence of Fokker-Planck type equations, we refer to [9, 3, 89] discussing several types of Fokker-Planck equation in different senses.

3.2 Brownian motion constructed by smooth orthonormal frames

The classical Eells-Elworth-Malliavin approach is an intrinsic approach utilizing only orthonormal frames for the tangent bundle, which is usually difficult to use for designing numerical schemes. An extrinsic approach using embedding was given in [43]. In this section, we provide a different yet more explicit intrinsic construction of Brownian motion, under the assumption of the existence of an orthonormal global smooth frame, i.e., for *parallelizable* manifolds, e.g., $\mathcal{S}_+^{n,p}$. The assumption of *Parallelizable* manifolds is quite restrictive since most manifolds such as S^2 are not parallelizable, see [60]. On the other hand, such an assumption can be relaxed by considering partition of unity, e.g., the approach in this section can still be used to get an intrinsic construction of Brownian motion on *non-parallelizable* manifolds like S^2, see Remark 19 in Section 3.2.2.

3.2.1 Notation

Our notation about Riemannian geometry follows most of those in [30, 1], and Einstein summation convention is used. All functions and vector fields, differential forms are assumed to be smooth unless otherwise stated. For a given Riemannian manifold $(\mathcal{M}, g)$ with dimension $\dim \mathcal{M} = d$, $T_x\mathcal{M}$ and $T_x^*\mathcal{M}$ denote its tangent space and cotangent space at $x \in \mathcal{M}$, $T\mathcal{M}$ and $T^*\mathcal{M}$ denote its tangent bundle and cotangent bundle, $\mathfrak{X}(T\mathcal{M})$ and $\mathfrak{X}(T^*\mathcal{M})$ denote the collection of all smooth vector fields and 1-form fields on $\mathcal{M}$. For simplicity, $\mathfrak{X}(T\mathcal{M})$ is just written as $\mathfrak{X}(\mathcal{M})$.

Although our main focus in this paper is $\mathcal{S}_+^{n,p}$, results in this section apply to a more general class of manifolds. To this end, in this section, we will use ***bold*** font for a point $\boldsymbol{X} \in \mathcal{M}$ or a vector field $\boldsymbol{v} \in \mathfrak{X}(\mathcal{M})$ to emphasize that they may not be necessarily related to $\mathcal{S}_+^{n,p}$. The directional derivative of function f along vector field $\boldsymbol{v}$ is denoted by $\boldsymbol{v}(f)$, which is also equal to the contraction of $\boldsymbol{v}$ and 1-form $\mathrm{d}f$ (written as $\boldsymbol{v}(\mathrm{d}f)$ or $\mathrm{d}f(\boldsymbol{v})$).

Given $\boldsymbol{u}, \boldsymbol{v} \in \mathfrak{X}(\mathcal{M})$, $\nabla_{\boldsymbol{u}}\boldsymbol{v}$ denotes the Levi-Civita connection. Notice that ∇ is also used to denote the *covariant differential* of a tensor ([30, Definition 5.7]), e.g., for $f \in C^\infty(\mathcal{M})$, $\nabla f = \mathrm{d}f \in \mathfrak{X}(T^*\mathcal{M})$, and for $\omega \in \mathfrak{X}(T^*\mathcal{M})$, $\nabla\omega$ is defined by $\nabla\omega(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}(\omega(\boldsymbol{v})) - \omega(\nabla_{\boldsymbol{u}}\boldsymbol{v}) \in \mathfrak{X}(T^*\mathcal{M} \otimes T^*\mathcal{M})$ for $\boldsymbol{u}, \boldsymbol{v} \in \mathfrak{X}(\mathcal{M})$.

Let $x : \mathcal{M} \to \Omega \subset \mathbb{R}^d$ be a coordinate chart. The *Riemannian gradient* operator $\mathrm{grad} : C^\infty(\mathcal{M}) \to \mathfrak{X}(\mathcal{M})$ is defined by

$$\mathrm{grad} f = g^{ij}\frac{\partial f}{\partial x^i}\left(\frac{\partial}{\partial x^j}\right), \qquad \forall f \in C^\infty(\mathcal{M}),$$

where g^{ij} defined by $g^{ij}g_{jk} = \delta^i_k$ is the inverse metric.

The divergence operator $\mathrm{div} : \mathfrak{X}(\mathcal{M}) \to C^\infty(\mathcal{M})$ maps a vector field $\boldsymbol{v} = v^i(\frac{\partial}{\partial x^i})$ to a function on $\mathcal{M}$:

$$\mathrm{div}\, \boldsymbol{v} = \frac{1}{\sqrt{\det g}}\frac{\partial}{\partial x^i}(\sqrt{\det g}\, v^i),$$

where $\det g = \det(g_{ij})$ is the determinant of matrix g_{ij}.

The Riemannian Hessian is defined by $\mathrm{Hess}\, f = \nabla(\mathrm{d}f) = \nabla^2 f$:

$$\mathrm{Hess}\, f = \nabla^2 f = \left(\frac{\partial^2 f}{\partial x^i \partial x^j} - \Gamma^k_{ij}\frac{\partial f}{\partial x^k}\right)\mathrm{d}x^i \otimes \mathrm{d}x^j,$$

where Γ^k_{ij} is Christoffel symbol. The Riemannian Hessian tensor is symmetric:

$$\mathrm{Hess}\, f(\boldsymbol{u}, \boldsymbol{v}) = g(\nabla_{\boldsymbol{u}}\mathrm{grad} f, \boldsymbol{v}) = g(\nabla_{\boldsymbol{v}}\mathrm{grad} f, \boldsymbol{u}), \quad \forall \boldsymbol{u}, \boldsymbol{v} \in \mathfrak{X}(\mathcal{M}).$$

The Riemannian Hessian satisfies ([1, Prop. 5.5.2]):

$$\operatorname{Hess} f(\boldsymbol{u}, \boldsymbol{v}) = \nabla^2 f(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}(\boldsymbol{v}f) - (\nabla_{\boldsymbol{u}}\boldsymbol{v})f, \quad \forall f \in C^\infty(\mathcal{M}), \forall \boldsymbol{u}, \boldsymbol{v} \in \mathfrak{X}(\mathcal{M}),$$

which implies

$$\nabla^2 f(\boldsymbol{v}, \boldsymbol{v}) = \boldsymbol{v}(\boldsymbol{v}f) - (\nabla_{\boldsymbol{v}}\boldsymbol{v})f, \quad \forall f \in C^\infty(\mathcal{M}), \forall \boldsymbol{v} \in \mathfrak{X}(\mathcal{M}). \tag{3.2.1}$$

The Laplace-Beltrami operator is defined by

$$\Delta_g f = \operatorname{div}(\operatorname{grad} f), \qquad \forall f \in C^\infty(\mathcal{M}).$$

Given an orthonormal basis $\{\boldsymbol{v}'_i\}_{i=1}^d \subset \mathfrak{X}(\mathcal{M})$, Δ_g is equal to the trace of ∇^2:

$$\Delta_g f = \sum_{i=1}^d \nabla^2 f(\boldsymbol{v}'_i, \boldsymbol{v}'_i), \quad \forall f \in C^\infty(\mathcal{M}). \tag{3.2.2}$$

Following definition of standard Brownian motion on manifolds [84, 43], we define

Definition 11. The Brownian motion with temperature $T = \frac{1}{\beta}$ on $(\mathcal{M}, g)$, denoted by $\boldsymbol{B}_t^{g,\beta}$, is a $\mathcal{M}$-valued random processes with the generator $\frac{1}{\beta}\Delta_g$ satisfying

$$f(\boldsymbol{B}_t^{g,\beta}) = f(\boldsymbol{B}_0^{g,\beta}) + \int_0^t \frac{1}{\beta}\Delta_g f(\boldsymbol{B}_s^{g,\beta}) \mathrm{d}s + \text{ some local martingale }, \forall f \in C^\infty(\mathcal{M}).$$

3.2.2 Brownian motion as a solution to SDE

Our first result is a simple way to construct Brownian motion on Riemannian manifolds using SDE, when an orthonormal basis of $\mathfrak{X}(\mathcal{M})$ is given. A basis of $\mathfrak{X}(\mathcal{M})$ is also called a *global smooth frame*, and a manifold with a global smooth frame is called *parallelizable* [60]. An orthonormal basis of $\mathfrak{X}(\mathcal{M})$ can be obtained by applying

Gram-Schmit procedure to a global smooth frame, see [29, Section 7.7].

Lemma 12. *For a parallelizable Riemannian manifold* $(\mathcal{M}, g)$ *with dimension* d*, let* $\{\boldsymbol{v}_i\}_{i=1}^d$ *be an orthonormal basis of* $\mathfrak{X}(\mathcal{M})$*. Let* $\boldsymbol{B}_t^{g,\beta}$ *be a Brownian motion on* $(\mathcal{M}, g)$ *with temperature* $T = \frac{1}{\beta}$*, then* $\boldsymbol{B}_t^{g,\beta}$ *is a solution to the following SDE:*

$$\mathrm{d}\boldsymbol{X}_t = \sum_{i=1}^{d} \big(\sqrt{\frac{2}{\beta}}\boldsymbol{v}_i(\boldsymbol{X}_t) \circ \mathrm{d}W_t^i - \frac{1}{\beta}\nabla_{\boldsymbol{v}_i}\boldsymbol{v}_i(\boldsymbol{X}_t)\mathrm{d}t\big), \tag{3.2.3}$$

where $\{W_t^i\}_{i=1}^d$ *are* d *independent* $\mathbb{R}$*-valued Wiener processes.*

Proof. By Definition 11, we only need to show that the SDE solution $\boldsymbol{X}_t$ satisfies

$$f(\boldsymbol{X}_t) - f(\boldsymbol{X}_0) = \int_0^t \frac{1}{\beta}\Delta_g f(\boldsymbol{X}_s)\mathrm{d}s + \text{some martingale}, \quad \forall f \in C^\infty(\mathcal{M}).$$

By [43, Definition 1.2.3], $f(\boldsymbol{X}_t)$ satisfies

$$\begin{aligned}
f(\boldsymbol{X}_t) - f(\boldsymbol{X}_0) &= \int_0^t \sum_{i=1}^{d} \big(\sqrt{\frac{2}{\beta}}\boldsymbol{v}_i f(\boldsymbol{X}_s) \circ \mathrm{d}W_s^i - \frac{1}{\beta}\nabla_{\boldsymbol{v}_i}\boldsymbol{v}_i f(\boldsymbol{X}_s)\mathrm{d}s\big) \\
&= \int_0^t \sum_{i=1}^{d} \big(\sqrt{\frac{2}{\beta}}\boldsymbol{v}_i f(\boldsymbol{X}_s)\mathrm{d}W_s^i + \frac{1}{\beta}\boldsymbol{v}_i(\boldsymbol{v}_i f)(\boldsymbol{X}_s)\mathrm{d}s - \frac{1}{\beta}\nabla_{\boldsymbol{v}_i}\boldsymbol{v}_i f(\boldsymbol{X}_s)\mathrm{d}s\big) \\
&= \int_0^t \sum_{i=1}^{d} \frac{1}{\beta}\nabla^2 f(\boldsymbol{v}_i, \boldsymbol{v}_i)(\boldsymbol{X}_s)\mathrm{d}s + \int_0^t \sum_{i=1}^{d} \sqrt{\frac{2}{\beta}}\boldsymbol{v}_i f(\boldsymbol{X}_s)\mathrm{d}W_s^i \\
&= \int_0^t \frac{1}{\beta}\Delta_g f(\boldsymbol{X}_s)\mathrm{d}s + \text{local martingale}
\end{aligned} \tag{3.2.4}$$

where we have used Strantonich-Itô conversion, (3.2.1) and (3.2.2).

□

Remark 17. The assumption of the manifold being parallelizable, e.g., existence of global smooth basis of $\mathfrak{X}(\mathcal{M})$, does not hold in general. For instance, the only parallelizable spheres are $\mathbb{S}^1, \mathbb{S}^3, \mathbb{S}^7$, see [60]. Thus Lemma 12 only applies to very special manifolds.

Remark 18. This result is consistent with the well known fact that Brownian motions are solutions to SDE in the form $\mathrm{d}X_t = V_\alpha(X_t) \circ \mathrm{d}W_t^\alpha + V_0(X_t)dt$, if the Laplace-Beltrami operator can be written in the form $\Delta_g = \frac{1}{2}\sum_\alpha V_\alpha^2 + V_0$, see [43]. The result in Lemma 12 simply means that the Laplace-Beltrami operator can be written in such a form using orthonormal basis of $\mathfrak{X}(\mathcal{M})$ on a parallelizable manifold.

Remark 19. The results in Lemma 12 can be extended to more general manifolds, by using partition of unity. See [30, Theorem 5.6] for existence of differentiable of partition of unity. Assume $\{(\rho_i, U_i)\}_{i\in\mathcal{I}}$ is a partition of unity of the manifold $\mathcal{M}$ and each $U_i \subset \mathcal{M}$ is parallelizable, Lemma 12 can be used to construct $\frac{1}{\beta}\rho_i\Delta_g$ on each $U_i \subset \mathcal{M}$. For example, the sphere $\mathbb{S}^2$ is not parallelizable, but we can have a partition of unity $(\rho_1, U_1), (\rho_2, U_2)$ where U_1, U_2 are $\mathbb{S}^2$ with north and south pole removed, then we can have an orthonormal basis $\{\boldsymbol{v}_{i,U_j}\}_{i=1}^2$ on U_j for each j. Define $\boldsymbol{u}_{i,j} = \sqrt{\rho_j}\boldsymbol{v}_{i,U_j}$ and consider the following SDE

$$\mathrm{d}\boldsymbol{X}_t = \sum_{i=1}^{2}\sum_{j=1}^{2}\Big(\sqrt{\frac{1}{\beta}}\boldsymbol{u}_{i,j}(\boldsymbol{X}_t) \circ \mathrm{d}W_t^{i,j} - \frac{1}{\beta}\nabla_{\boldsymbol{u}_{i,j}}\boldsymbol{u}_{i,j}(\boldsymbol{X}_t)\mathrm{d}t\Big)\,, \tag{3.2.5}$$

where $\{W_t^{i,j}\}_{i,j=1}^2$ are 2×2 of independent $\mathbb{R}$ valued Wiener processes. Because ρ_i vanishes outside of U_i, $\boldsymbol{u}_{i,j}$ is well defined on the whole $\mathbb{S}^2$, so is the SDE (3.2.5). Following the proof of Lemma 12, the solution to (3.2.5) is the Brownian motion on $\mathbb{S}^2$.

As explained in Remark (19) **above, partition of unity will not essentially affect the key results in following contents. But for simplicity, we will consider only the parallelizable manifolds.**

As Lemma 12 shows that we are able to construct $\boldsymbol{B}_t^{g,\beta}$ using SDE (3.2.3), formally we also use $\mathrm{d}\boldsymbol{B}_t^{g,\beta}$ to denote the R.H.S of SDE (3.2.3) in RLE (3.1.2), where all vector fields are evaluated at $\boldsymbol{X}_t$ instead of $\boldsymbol{B}_t^{g,\beta}$, i.e. by RLE (3.1.2), we actually mean the following SDE

$$\mathrm{d}\boldsymbol{X}_t = -\mathrm{grad}\mathcal{E}(\boldsymbol{X}_t)\mathrm{d}t + \sum_{i=1}^{n}(\sqrt{\frac{2}{\beta}}\boldsymbol{v}_i(\boldsymbol{X}_t) \circ \mathrm{d}W_t^i - \frac{1}{\beta}\nabla_{\boldsymbol{v}_i}\boldsymbol{v}_i(\boldsymbol{X}_t)\mathrm{d}t) \tag{3.2.6}$$

where $\{\boldsymbol{v}_i\}_{i=1}^n$ is an orthonormal basis of $\mathfrak{X}(\mathcal{M})$.

3.2.3 Brownian motion on an embedded submanifold

Let $(\mathcal{M}, g)$ with dimension d be a Riemannian submanifold of $(\widetilde{\mathcal{M}}, \tilde{g})$ of dimension $\tilde{d}$ with an isometric inclusion map $i : \mathcal{M} \to \widetilde{\mathcal{M}}$, dimensions $d, \tilde{d}$, and Levi-Civita connections $\nabla, \widetilde{\nabla}$ respectively. For convenience, we regard $\boldsymbol{X} \in \mathcal{M}$ the same as its image $i(\boldsymbol{X}) \in \widetilde{\mathcal{M}}$, we regard $\boldsymbol{v} \in T_{\boldsymbol{X}}\mathcal{M}$ the same as its image $\mathrm{d}i(\boldsymbol{v}) \in T_{i(\boldsymbol{X})}\widetilde{\mathcal{M}}$. For an isometric immersion (see [30, 55, 52]), we have

$$T_{i(\boldsymbol{X})}\widetilde{\mathcal{M}} = T_{\boldsymbol{X}}\mathcal{M} \oplus (T_{\boldsymbol{X}}\mathcal{M})^{\perp}$$

and the Levi-Civita connections satisfy :

$$\nabla_{\boldsymbol{u}}\boldsymbol{v} = \widetilde{\nabla}_{\boldsymbol{u}}\boldsymbol{v} - (\widetilde{\nabla}_{\boldsymbol{u}}\boldsymbol{v})^{\perp}, \tag{3.2.7}$$

where $\mathbf{u}, \mathbf{v}$ are local smooth vector fields on $\mathcal{M}$ and $(\cdot)^{\perp}$ is the orthogonal projection to $(T\mathcal{M})^{\perp}$. The second fundamental form $\boldsymbol{I\!I} : T\mathcal{M} \otimes T\mathcal{M} \to (T\mathcal{M})^{\perp}$ is defined by (see [30, 55]):

$$\boldsymbol{I\!I}(\boldsymbol{u}, \boldsymbol{v}) = (\widetilde{\nabla}_{\boldsymbol{u}}\boldsymbol{v})^{\perp}\,, \tag{3.2.8}$$

which is a bilinear map and does not depend on the derivatives of $\boldsymbol{u}, \boldsymbol{v}$.

At each $\boldsymbol{X} \in \mathcal{M}$, for each $\boldsymbol{\eta} \in (T_{\boldsymbol{X}}\mathcal{M})^{\perp}$, $\boldsymbol{I\!I}$ is associated to a linear self-adjoint map $S_{\boldsymbol{\eta}} : T_{\boldsymbol{X}}\mathcal{M} \to T_{\boldsymbol{X}}\mathcal{M}$:

$$\tilde{g}(S_{\boldsymbol{\eta}}(\boldsymbol{u}), \boldsymbol{v}) = \tilde{g}(\boldsymbol{I\!I}(\boldsymbol{u}, \boldsymbol{v}), \boldsymbol{\eta}) \tag{3.2.9}$$

for all $\boldsymbol{u}, \boldsymbol{v} \in T_{\boldsymbol{X}}\mathcal{M}$. The trace of $S_{\boldsymbol{\eta}}$ is well defined for each $\boldsymbol{\eta}$ and is linear in $\boldsymbol{\eta}$. The *mean curvature vector* [30, 52] of $\mathcal{M}$ at $\boldsymbol{X}$ is defined as $\boldsymbol{H} \in (T_{\boldsymbol{X}}\mathcal{M})^{\perp}$ satisfying

$$\tilde{g}(\boldsymbol{H}, \boldsymbol{\eta}) = \frac{1}{d}\mathrm{Tr}\, S_{\boldsymbol{\eta}}, \quad \text{for all } \boldsymbol{\eta} \in (T_{\boldsymbol{X}}\mathcal{M})^{\perp}. \tag{3.2.10}$$

With an orthonormal basis $\{\boldsymbol{v}_i\}_{i=1}^{d} \subset \mathfrak{X}(\mathcal{M})$ locally defined in the neighborhood of $\boldsymbol{X}$, we can easily find the trace of $S_{\boldsymbol{\eta}}$. By (3.2.8), (3.2.9) and (3.2.10), the mean curvature vector has the expression

$$\boldsymbol{H}(\boldsymbol{X}) = \frac{1}{d}\sum_{i=1}^{d}(\widetilde{\nabla}_{\boldsymbol{v}_i}\boldsymbol{v}_i)^{\perp}\,. \tag{3.2.11}$$

We first state a result connecting Brownian motion with the mean curvature vector, which was first discussed by Lewis in [90, 61] for submanifolds in Euclidean space, see also [82, Chapter 5]. The mean curvature vector was also pointed out by Stroock in [84, Chapter 4] and [85, Section 5] for a general embedded manifold. In the following theorem, for a parallelizable manifold submanifold, we are able to give a more explicit characterization of relation between Brownian motion and mean curvature vector.

Theorem 12. *Let $(\mathcal{M}, g)$ be an Riemannian submanifold of $(\widetilde{\mathcal{M}}, \tilde{g})$, with the isometric immersion $i : \mathcal{M} \to \widetilde{\mathcal{M}}$, dimensions $d, \tilde{d}$, and Levi-Civita connections $\nabla, \widetilde{\nabla}$ respectively. Assume $\mathcal{M}$ is parallelizable, with an orthonormal basis $\{\boldsymbol{v}_i\}_{i=1}^{d}$ of $\mathfrak{X}(\mathcal{M})$, then the Brownian motion $\boldsymbol{B}_t^{g,\beta}$ on $\mathcal{M}$ is the solution to SDE* (3.2.3)*. Let $\boldsymbol{Z}_t = i(\boldsymbol{B}_t^{g,\beta})$*

be its image in $\widetilde{\mathcal{M}}$, then for any $\tilde{f} \in C^\infty(\widetilde{\mathcal{M}})$, the evolution of $\tilde{f}(\boldsymbol{Z}_t)$ is given by

$$\mathrm{d}\tilde{f}(\boldsymbol{Z}_t) = \frac{1}{\beta}\sum_{i=1}^{d} \widetilde{\nabla}^2 \tilde{f}(\boldsymbol{v}_i, \boldsymbol{v}_i)\mathrm{d}t + \frac{d}{\beta}\boldsymbol{H}\tilde{f}(\boldsymbol{Z}_t)\mathrm{d}t + \sum_{i=1}^{d}\sqrt{\frac{2}{\beta}}\boldsymbol{v}_i\tilde{f}(\boldsymbol{Z}_t)\mathrm{d}W_t^i, \tag{3.2.12}$$

where $\sum_{i=1}^{d} \widetilde{\nabla}^2 \tilde{f}(\boldsymbol{v}_i, \boldsymbol{v}_i)$ is the trace of $\widetilde{\nabla}^2 \tilde{f}$ in the subspace $T\mathcal{M} \subset T\widetilde{\mathcal{M}}$.

Remark 20.

$$\mathrm{d}\tilde{f}(\boldsymbol{Z}_t) = \frac{1}{\beta}\mathrm{Tr}|_{T\mathcal{M}}\tilde{\nabla}^2\tilde{f}\mathrm{d}t + \frac{d}{\beta}\boldsymbol{H}\tilde{f}(\boldsymbol{Z}_t)\mathrm{d}t + \text{local martingale}, \tag{3.2.13}$$

Proof. Let $f = \tilde{f} \circ i$ be the restriction of $\tilde{f}$ on $\mathcal{M}$, then $f(\boldsymbol{B}_t^{g,\beta}) = \tilde{f}(\boldsymbol{Z}_t)$. By Lemma 12, the evolution of $f(\boldsymbol{B}_t^{g,\beta})$ can be described as

$$\mathrm{d}f(\boldsymbol{B}_t^{g,\beta}) = \sum_{i=1}^{d}\left(\sqrt{\frac{2}{\beta}}\boldsymbol{v}_i f(\boldsymbol{B}_t^{g,\beta}) \circ \mathrm{d}W_t^i - \frac{1}{\beta}\nabla_{\boldsymbol{v}_i}\boldsymbol{v}_i f(\boldsymbol{B}_t^{g,\beta})\mathrm{d}t\right), \tag{3.2.14}$$

and we just need to express them in terms of $\tilde{f}$ and its derivatives in $\widetilde{\mathcal{M}}$. By regarding $\boldsymbol{v}_i$ as a vector field also in $\mathfrak{X}(\widetilde{\mathcal{M}})$, we have

$$\boldsymbol{v}_i\tilde{f}(\boldsymbol{Z}_t) = \boldsymbol{v}_i f(\boldsymbol{B}_t^{g,\beta}). \tag{3.2.15}$$

The Levi-Civita connections for an isometric immersion gives (3.2.7) thus

$$\nabla_{\boldsymbol{v}_i}\boldsymbol{v}_i = \widetilde{\nabla}_{\boldsymbol{v}_i}\boldsymbol{v}_i - (\widetilde{\nabla}_{\boldsymbol{v}_i}\boldsymbol{v}_i)^\perp. \tag{3.2.16}$$

Thus we have

$$
\begin{aligned}
&\mathrm{d}\tilde{f}(\boldsymbol{Z}_t) = \mathrm{d}f(\boldsymbol{B}_t^{g,\beta}) \qquad (3.2.17)\\
=&\sum_{i=1}^{d}\Big(\sqrt{\frac{2}{\beta}}\boldsymbol{v}_i\tilde{f}(\boldsymbol{Z}_t)\circ \mathrm{d}W_t^i - \frac{1}{\beta}\widetilde{\nabla}_{\boldsymbol{v}_i}\boldsymbol{v}_i\tilde{f}(\boldsymbol{Z}_t)\mathrm{d}t\Big) + \frac{1}{\beta}\sum_{i=1}^{d}(\widetilde{\nabla}_{\boldsymbol{v}_i}\boldsymbol{v}_i)^{\perp}\tilde{f}(\boldsymbol{Z}_t)\mathrm{d}t\\
=&\sum_{i=1}^{d}\Big(\sqrt{\frac{2}{\beta}}\boldsymbol{v}_i\tilde{f}(\boldsymbol{Z}_t)\mathrm{d}W_t^i + \frac{1}{\beta}\widetilde{\nabla}^2\tilde{f}(\boldsymbol{v}_i,\boldsymbol{v}_i)(\boldsymbol{Z}_t)\mathrm{d}t\Big) + \frac{1}{\beta}d\boldsymbol{H}\tilde{f}(\boldsymbol{Z}_t)\mathrm{d}t,
\end{aligned}
$$

where we have used Strantonich-Itô conversion and also (3.2.1) in the last step, similar to the derivation in (3.2.4). □

3.2.4 Brownian motion via Riemannian submersion

A surjective submersion $\pi : (\widetilde{\mathcal{M}}, \tilde{g}) \to (\mathcal{M}, g)$ is Riemannian if $\mathrm{d}\pi$ is a linear isometry from $\mathrm{Ker}(\mathrm{d}\pi)^{\perp}$ onto $T_{\boldsymbol{X}}\mathcal{M}$ for all $\boldsymbol{X} \in \mathcal{M}$. Properties and facts about Riemannian submersion that we will use can be found in [76]. For each $\boldsymbol{Z} \in \mathcal{M}$, $\pi^{-1}(\boldsymbol{Z}) \subset \widetilde{\mathcal{M}}$ is also a Riemannian submanifold of $\widetilde{\mathcal{M}}$ and it is also called a *fiber*.

Next, we state the result that the Brownian motions via Riemannian submersion is also related to the mean curvature of each fiber. This result is first proved in [49].

Theorem 13. *Let $\pi : (\widetilde{\mathcal{M}}, \tilde{g}) \to (\mathcal{M}, g)$ be a Riemannian submersion. Let $\boldsymbol{H}(\boldsymbol{X})$ be the mean curvature normal vector at $\boldsymbol{X}$ of the fiber $\pi^{-1}(\pi(\boldsymbol{X})) \subset \widetilde{\mathcal{M}}$. If $\widetilde{\mathcal{M}}$ and $\mathcal{M}$ have dimensions $\tilde{d}$ and d, then each fiber has dimension $\tilde{d} - d$. For any fixed $\boldsymbol{Z} = \pi(\boldsymbol{X})$, assume that $\boldsymbol{h} = \mathrm{d}\pi(\boldsymbol{H}(\boldsymbol{X}))$ is independent of different $\boldsymbol{X} \in \pi^{-1}(\boldsymbol{Z})$, so that $\boldsymbol{h}(\boldsymbol{Z})$ is a vector field in $\mathfrak{X}(\mathcal{M})$. Let $B_t^{g,\beta}$ be the Brownian motin on $(\mathcal{M}, g)$, $B_t^{\tilde{g},\beta}$ be the Brownian motin on $(\widetilde{\mathcal{M}}, \tilde{g})$ and $\boldsymbol{Z}_t = \pi(\boldsymbol{B}_t^{\tilde{g},\beta})$, then $\boldsymbol{Z}_t$ solves the following SDE*

$$
\mathrm{d}\boldsymbol{Z}_t = -\frac{\tilde{d}-d}{\beta}\boldsymbol{h}(\boldsymbol{Z}_t)\mathrm{d}t + \mathrm{d}\boldsymbol{B}_t^{g,\beta}, \qquad (3.2.18)
$$

or equivalently,

$$\mathrm{d}f(\boldsymbol{Z}_t) = -\frac{\tilde{d}-d}{\beta}\boldsymbol{h}f(\boldsymbol{Z}_t)\mathrm{d}t + \frac{1}{\beta}\Delta_g f(\boldsymbol{Z}_t)\mathrm{d}t + \textit{ local martingale}\,, \quad \forall f \in C^\infty(\mathcal{M}). \tag{3.2.19}$$

Proof. For each $\boldsymbol{X} \in \widetilde{\mathcal{M}}, \boldsymbol{Z} = \pi(\boldsymbol{X}) \in \mathcal{M}$, $\tilde{f} = f \circ \pi \in C^\infty(\widetilde{\mathcal{M}})$ is a constant on $\pi^{-1}(\boldsymbol{Z})$, we have

$$\boldsymbol{v}_i\tilde{f} = 0 \qquad i = 1, 2, ..., d' - d\,, \tag{3.2.20}$$

$$\widetilde{\nabla}_{\boldsymbol{v}_i}\boldsymbol{v}_i\tilde{f} = (\widetilde{\nabla}_{\boldsymbol{v}_i}\boldsymbol{v}_i)^\perp\tilde{f} \qquad i = 1, 2, ..., d' - d\,. \tag{3.2.21}$$

Let $\tilde{f} = f \circ \pi \in C^\infty(\widetilde{\mathcal{M}})$, then $f(\boldsymbol{Z}_t) = \tilde{f}(\boldsymbol{B}_t^{\tilde{g},\beta})$ and

$$\mathrm{d}f(\boldsymbol{Z}_t) = \mathrm{d}\tilde{f}(\boldsymbol{B}_t^{\tilde{g},\beta}) = \frac{1}{\beta}\Delta_{\tilde{g}}\tilde{f}(\boldsymbol{B}_t^{\tilde{g},\beta})\mathrm{d}t + \text{ local martingale,} \tag{3.2.22}$$

thus we only need to prove the following identity for $\boldsymbol{X} \in \widetilde{\mathcal{M}}$ and $\boldsymbol{Z} = \pi(\boldsymbol{X})$:

$$\Delta_{\tilde{g}}\tilde{f}(\boldsymbol{X}) = \Delta_g f(\boldsymbol{Z}) - (\tilde{d} - d)\boldsymbol{h}f(\boldsymbol{Z})\,. \tag{3.2.23}$$

To this end, consider a smooth curve $\tilde{\gamma} : [-1, 1] \to \widetilde{\mathcal{M}}$ with $\tilde{\gamma}(0) = \boldsymbol{X}$, let $\gamma(t) = \pi(\tilde{\gamma}(t))$. Let $\tilde{\boldsymbol{V}}$ and $\boldsymbol{V}$ be the vector fields such that $\dot{\tilde{\gamma}}(t) = \tilde{\boldsymbol{V}}(\tilde{\gamma}(t))$ and $\dot{\gamma}(t) = \boldsymbol{V}(\gamma(t))$. Take derivative of $f(\gamma(t)) = \tilde{f}(\tilde{\gamma}(t))$ at $t = 0$, we have

$$\boldsymbol{V}f(\boldsymbol{Z}) = \tilde{\boldsymbol{V}}\tilde{f}(\boldsymbol{X}). \tag{3.2.24}$$

Take the second order derivative, $\boldsymbol{V}(\boldsymbol{V}f)(\boldsymbol{Z}) = \tilde{\boldsymbol{V}}(\tilde{\boldsymbol{V}}\tilde{f})(\boldsymbol{X})$. With (3.2.1), we have

$$\nabla_{\boldsymbol{V}}\boldsymbol{V}f(\boldsymbol{Z}) + \nabla^2 f(\boldsymbol{V}, \boldsymbol{V})(\boldsymbol{Z}) = \tilde{\nabla}_{\tilde{\boldsymbol{V}}}\tilde{\boldsymbol{V}}\tilde{f}(\boldsymbol{X}) + \tilde{\nabla}^2\tilde{f}(\tilde{\boldsymbol{V}}, \tilde{\boldsymbol{V}})(\boldsymbol{X}). \tag{3.2.25}$$

Next, we consider orthononormal vector fields $E = \{\tilde{\boldsymbol{V}}_i\}_{i=1}^{\tilde{d}}$ locally defined on $\widetilde{\mathcal{M}}$ in a neighborhood U around $\boldsymbol{X}$, which satisfies that $\{\tilde{\boldsymbol{V}}_i(\boldsymbol{Y})\}_{i=1}^{\tilde{d}}$ is an orthonormal basis of $T_{\boldsymbol{Y}}\widetilde{\mathcal{M}}$ and $E_1 = \{\tilde{\boldsymbol{V}}_i(\boldsymbol{Y})\}_{i=1}^{d}$ is an orthonormal basis of $\mathrm{Ker}(\mathrm{d}\pi)^{\perp}$ at $\boldsymbol{Y}$, for any $\boldsymbol{Y} \in U \subset \widetilde{\mathcal{M}}$. Such an orthonormal basis can be obtained by Gram-Schmit procedure to smooth vector fields [29, Section 7.7]. Then $E_2 = \{\tilde{\boldsymbol{V}}_i\}_{i=d+1}^{\tilde{d}}$ is an orthonormal basis of $\mathrm{Ker}(\mathrm{d}\pi)$. In other words, E_1 contains orthonormal *horizontal* vector fields and E_2 contains orthonormal *vertical* vector fields.

For $\tilde{\boldsymbol{V}} \in E_2$, $\boldsymbol{V} = \mathrm{d}\pi(\tilde{\boldsymbol{V}}) = 0$, so the L.H.S is zero. Take summation we have

$$0 = \sum_{\tilde{\boldsymbol{V}} \in E_2} \Big(\tilde{\nabla}_{\tilde{\boldsymbol{V}}}\tilde{\boldsymbol{V}}\tilde{f}(\boldsymbol{X}) + \tilde{\nabla}^2\tilde{f}(\tilde{\boldsymbol{V}},\tilde{\boldsymbol{V}})(\boldsymbol{X})\Big). \tag{3.2.26}$$

For $\tilde{\boldsymbol{V}} \in E_1$, we have $\nabla_{\boldsymbol{V}}\boldsymbol{V} = \mathrm{d}\pi(\tilde{\nabla}_{\tilde{\boldsymbol{V}}}\tilde{\boldsymbol{V}})$, therefore

$$\sum_{\boldsymbol{V} \in E} \nabla^2 f(\boldsymbol{V},\boldsymbol{V})(\boldsymbol{Z}) = \sum_{\tilde{\boldsymbol{V}} \in E_1} \tilde{\nabla}^2\tilde{f}(\tilde{\boldsymbol{V}},\tilde{\boldsymbol{V}})(\boldsymbol{X}). \tag{3.2.27}$$

thus,

$$\Delta_g f(\boldsymbol{Z}) = \sum_{\tilde{\boldsymbol{V}} \in E_2} \tilde{\nabla}_{\tilde{\boldsymbol{V}}}\tilde{\boldsymbol{V}}\tilde{f}(\boldsymbol{X}) + \Delta_{\tilde{g}}\tilde{f}(\boldsymbol{X}) \tag{3.2.28}$$

Since $\tilde{f}$ is a constant over $\pi^{-1}(\boldsymbol{Z})$, only the part of $\tilde{\nabla}_{\tilde{\boldsymbol{V}}}\tilde{\boldsymbol{V}}$ in $\mathrm{Ker}(\pi_*)^{\perp} = T(\pi^{-1}(\boldsymbol{Z}))^{\perp}$ makes non-zero contribution, thus

$$\begin{aligned}
&\sum_{\tilde{\boldsymbol{V}} \in E_2} \tilde{\nabla}_{\tilde{\boldsymbol{V}}}\tilde{\boldsymbol{V}}\tilde{f}(\boldsymbol{Y}) \\
=&\sum_{\tilde{\boldsymbol{V}} \in E_2} \big(\tilde{\nabla}_{\tilde{\boldsymbol{V}}}\tilde{\boldsymbol{V}}\big)^{\perp}\tilde{f}(\boldsymbol{Y}) \\
=&\boldsymbol{H}\tilde{f}(\boldsymbol{X})
\end{aligned} \tag{3.2.29}$$

By assumption on $\boldsymbol{H}$, $\boldsymbol{H}\tilde{f}(\boldsymbol{X}) = \boldsymbol{h}f(\boldsymbol{Z})$, so we are done. □

Remark 21. In the settings above, it is asserted in [78] that $-\boldsymbol{h} = \operatorname{grad} S$, $S(\boldsymbol{Z}) = \log \operatorname{vol}(\pi^{-1}(\boldsymbol{Z}))$, where grad is with respect to metric g, we conclude that $\boldsymbol{Z}_t$ solves the following RLE

$$\mathrm{d}\boldsymbol{Z}_t = \operatorname{grad} S(\boldsymbol{Z}_t)\mathrm{d}t + \mathrm{d}\boldsymbol{B}_t^{g,\beta}\,. \tag{3.2.30}$$

3.3 Explicit Riemannian Langevin equation on $\mathcal{S}_+^{n,p}$

In the following sections, we provide 2 RLEs in explicit form on the manifold $\mathcal{S}_+^{n,p}$ equipped with different metrics.

3.3.1 Embedded metric case

$\mathrm{Sym}(n) = \{X \in \mathbb{R}^{n\times n} | X = X^T\}$ could be equipped with Euclidean metric $g(U,V) = \mathrm{Tr}(UV^T)$ which induces a metric g_E on the submanifold $\mathcal{S}_+^{n,p}$, defined by

$$g_{E,X}(U,V) = \mathrm{Tr}(UV^T) \qquad\qquad X \in \mathcal{S}_+^{n,p},\, U,V \in T_X\mathcal{S}_+^{n,p} \tag{3.3.1}$$

We would like an explicit form of SDE (3.1.2) describing the evolution of the entries of the matrix $\{(X_t)_{ij}\}_{i,j=1}^n$, which will benefit numerical implementation. For g_E, we have the following theorem:

Theorem 14. *Let* $X_t = Q_t\Lambda_t Q_t^T$, $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_p, 0, ..., 0)$*, then RLE* (3.1.2) *takes the following form*

$$\mathrm{d}X_t = P_{X_t}(-\mathcal{D}\mathcal{E}(X_t))\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}W_t^{n,p,X_t} + \frac{1}{\beta}H(X_t)\mathrm{d}t \tag{3.3.2}$$

where $\mathcal{D}\mathcal{E}$ is the gradient of $\mathcal{E}$ under Euclidean metric of $\mathrm{Sym}(n)$

$$\mathcal{D}\mathcal{E} = \begin{bmatrix} \frac{\partial \mathcal{E}}{\partial X_{11}} & \frac{\partial \mathcal{E}}{\partial X_{12}} & \cdots & \frac{\partial \mathcal{E}}{\partial X_{1n}} \\ \frac{\partial \mathcal{E}}{\partial X_{21}} & \frac{\partial \mathcal{E}}{\partial X_{22}} & \cdots & \frac{\partial \mathcal{E}}{\partial X_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathcal{E}}{\partial X_{n1}} & \frac{\partial \mathcal{E}}{\partial X_{n2}} & \cdots & \frac{\partial \mathcal{E}}{\partial X_{nn}} \end{bmatrix} \tag{3.3.3}$$

P_{X_t} is the projection operator $T_{X_t}\mathrm{Sym}(n) \to T_{X_t}\mathcal{S}_+^{n,p}$,

$$P_{X_t}(Q_t \begin{bmatrix} A_{p\times p} & B_{p\times(n-p)} \\ (B_{p\times(n-p)})^T & C_{(n-p)\times(n-p)} \end{bmatrix} Q_t^T) = Q_t \begin{bmatrix} A_{p\times p} & B_{p\times(n-p)} \\ (B_{p\times(n-p)})^T & 0_{(n-p)\times(n-p)} \end{bmatrix} Q_t^T \tag{3.3.4}$$

$\mathrm{d}W_t^{n,p,X_t}$ is diffusion term in $T_{X_t}\mathcal{S}_+^{n,p}$, and $H(X_t)$ is the mean curvature correction.

$$\begin{aligned}
\mathrm{d}W_t^{n,p,X_t} =& Q_t \begin{bmatrix}
\mathrm{d}W_t^{1,1} & \cdots & \frac{1}{\sqrt{2}}\mathrm{d}W_t^{1,p} & \frac{1}{\sqrt{2}}\mathrm{d}W_t^{1,p+1} & \cdots & \frac{1}{\sqrt{2}}\mathrm{d}W_t^{1,n} \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\frac{1}{\sqrt{2}}\mathrm{d}W_t^{1,p} & \cdots & \mathrm{d}W_t^{p,p} & \frac{1}{\sqrt{2}}\mathrm{d}W_t^{p,p+1} & \cdots & \frac{1}{\sqrt{2}}\mathrm{d}W_t^{p,n} \\
\frac{1}{\sqrt{2}}\mathrm{d}W_t^{1,p+1} & \cdots & \frac{1}{\sqrt{2}}\mathrm{d}W_t^{p,p+1} & 0 & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\frac{1}{\sqrt{2}}\mathrm{d}W_t^{1,n} & \cdots & \frac{1}{\sqrt{2}}\mathrm{d}W_t^{p,n} & 0 & \cdots & 0
\end{bmatrix} Q_t^T \\
H(X_t) =& Q_t \begin{bmatrix} 0_{p\times p} & 0_{p\times(n-p)} \\ 0_{(n-p)\times p} & \sum_{i=1}^p \frac{1}{\lambda_i} I_{n-p} \end{bmatrix} Q_t^T
\end{aligned} \tag{3.3.5}$$

where $\{W_t^{i,j}\}_{1\le i\le p, i+1\le j\le n}$ are $np + \frac{p(p+1)}{2}$ independent standard Brownian motions in $\mathbb{R}$.

Proof of theorem 14. $(\mathcal{S}_+^{n,p}, g_E)$ is an embedded submanifold of $(\mathrm{Sym}(n), \mathrm{Tr}(\mathrm{d}X\mathrm{d}X^T))$, so corollary 12 applies. As a Euclidean space, the covariant derivative $\widetilde{\nabla} = \mathcal{D}$ is just the usual derivative with respect to X. For the evolution of coordinate X_t, notice that each function X_{ij} have vanishing Hessian $\widetilde{\nabla}^2 X_{ij} = \mathcal{D}^2 X_{ij} = 0$, so we only need to construct an orthonormal basis of $T\mathcal{S}_+^{n,p}$ and evaluate the mean curvature $\boldsymbol{H}$ of $\mathcal{S}_+^{n,p}$. For RLE (3.1.2), The drift term $\mathrm{grad}\,\mathcal{E}$ is just give by the projection of $\mathcal{D}\mathcal{E}$ to $T\mathcal{S}_+^{n,p}$.

The tangent space $T\mathcal{S}_+^{n,p}$ has been well studied in literature, see [96] for reference.

Here we just introduce an orthonormal basis of $(T\mathcal{S}_+^{n,p})$. For $X \in \mathcal{S}_+^{n,p}$, consider its diagonalization $X = Q\Lambda Q^T$, $Q \in \mathcal{O}_n$, $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_p, 0, ..., 0)$, $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$. We define the desired vector fields $\boldsymbol{v}_i$ by giving the directional derivatives $\boldsymbol{v}_i(X)$

$$
\begin{aligned}
&\boldsymbol{v}_i(X) = Qe_ie_i^TQ^T && 1 \le i \le p \\
&\boldsymbol{v}_{ij}(X) = Q\frac{e_ie_j^T + e_je_i^T}{\sqrt{2}}Q^T && 1 \le i \le p, i+1 \le j \le n
\end{aligned} \tag{3.3.6}
$$

where e_i is the i-th unit column vector.

Equivalently, we could define these vector fields using their action on Q and Λ:

$$
\begin{aligned}
&\boldsymbol{v}_i(\Lambda) = e_ie_i^T\ , \quad \boldsymbol{v}_i(Q) = 0 && 1 \le i \le n \\
&\boldsymbol{v}_{ij}(\Lambda) = 0\ , \quad \boldsymbol{v}_{ij}(Q) = Q\frac{-e_ie_j^T + e_je_i^T}{\sqrt{2}(\lambda_i - \lambda_j)} && 1 \le i \le p, i+1 \le j \le p \\
&\boldsymbol{v}_{ij}(\Lambda) = 0\ , \quad \boldsymbol{v}_{ij}(Q) = Q\frac{-e_ie_j^T + e_je_i^T}{\sqrt{2}\lambda_i} && 1 \le i \le p, p+1 \le j \le n
\end{aligned} \tag{3.3.7}
$$

Let $\mathcal{I} = \{1, ..., p\} \cup \{(ij) | 1 \le i \le p, i+1 \le j \le n\}$, we use $\{\boldsymbol{v}_I\}_{I \in \mathcal{I}}$ to denote this orthonormal basis.

To compute the mean curvature, we should compute $(\mathcal{D}_{\boldsymbol{v}_I}\boldsymbol{v}_I)^\perp$ for each $I \in \mathcal{I}$, where the normal space $N\mathcal{S}_+^{n,p}$ is known to be

$$
N_{Q\Lambda Q^T}\mathcal{S}_+^{n,p} = \big\{Q \begin{pmatrix} 0_{p\times p} & 0_{p\times(n-p)} \\ 0_{(n-p)\times p} & Y_{(n-p)\times(n-p)} \end{pmatrix} Q^T \big| Y \in \mathrm{Sym}(n-p)\big\} \tag{3.3.8}
$$

Still benefited from the vanishing Hessian of coordinate X, we have

$$
\boldsymbol{v}_I(\boldsymbol{v}_I(X)) = \mathcal{D}^2X(\boldsymbol{v}_I, \boldsymbol{v}_I) + \mathcal{D}_{\boldsymbol{v}_I}\boldsymbol{v}_I(X) = \mathcal{D}_{\boldsymbol{v}_I}\boldsymbol{v}_I(X) \tag{3.3.9}
$$

so

$$\begin{aligned}\mathcal{D}_{\boldsymbol{v}_I}\boldsymbol{v}_I(X) =&\boldsymbol{v}_I(\boldsymbol{v}_I(X)) = \boldsymbol{v}_I(\boldsymbol{v}_I(Q\Lambda Q^T)) = \boldsymbol{v}_I(Q(\boldsymbol{v}_I(\Lambda) + [\omega_I, \Lambda])Q^T)\\ =&Q(\boldsymbol{v}_I(\boldsymbol{v}_I(\Lambda)) + [\boldsymbol{v}_I(\omega_I), \Lambda] + [\omega_I, \boldsymbol{v}_I(\Lambda)] + [\omega_I, [\omega_I, \Lambda]])Q^T\\ =&Q[\omega_I, [\omega_I, \Lambda]]Q^T \end{aligned} \tag{3.3.10}$$

where $\omega_I = Q^{-1}\boldsymbol{v}_I(Q)$.

From some computation, we find

$$\begin{aligned} &(\mathcal{D}_{\boldsymbol{v}_i}\boldsymbol{v}_i)^\perp = 0 && i = 1, ..., n\\ &(\mathcal{D}_{\boldsymbol{v}_{ij}}\boldsymbol{v}_{ij})^\perp = 0 && 1 \le i, j \le p\\ &(\mathcal{D}_{\boldsymbol{v}_{ij}}\boldsymbol{v}_{ij})^\perp = Q\frac{e_j e_j^T}{\lambda_i}Q^T && 1 \le i \le p, p+1 \le j \le n \end{aligned} \tag{3.3.11}$$

and mean curvature

$$\boldsymbol{H}(Q\Lambda Q^T) = (\sum_{i=1}^{p} \frac{1}{\lambda_i})Q \begin{pmatrix} 0_{p\times p} & 0_{p\times(n-p)} \\ 0_{(n-p)\times p} & I_{(n-p)\times(n-p)} \end{pmatrix} Q^T \tag{3.3.12}$$

Now we can write down the evolution of X_t when X_t solves RLE (3.1.2), with each term explicitly known to us:

$$\mathrm{d}X_t = - P_{X_t}(\mathcal{D}\mathcal{E}(X_t))\mathrm{d}t + \frac{1}{\beta}\boldsymbol{H}(X_t)\mathrm{d}t + \sum_{I\in\mathcal{I}} \sqrt{\frac{2}{\beta}}\boldsymbol{v}_I(X_t)\mathrm{d}W_t^I \tag{3.3.13}$$

□

3.3.2 Bures-Wasserstein metric case

The manifold $\mathcal{S}_+^{n,p}$ can also be viewed as a quotient manifold $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$, for which the noncompact Stiefel manifold $\mathbb{R}_*^{n\times p}$ is called the *total space*. Denote the natural projection as

$$\tilde{\pi} : \mathbb{R}_*^{n\times p} \to \mathbb{R}_*^{n\times p}/\mathcal{O}_p.$$

For any $Y \in \mathbb{R}_*^{n\times p}$, the equivalence class containing Y is

$$[Y] = \tilde{\pi}^{-1}(\tilde{\pi}(Y)) = \{YO|O \in \mathcal{O}_p\},$$

which is an embedded submanifold of $\mathbb{R}_*^{n\times p}$([1, Prop. 3.4.4]). The tangent space of $[Y]$ at Y is therefore a subspace of $T_Y\mathbb{R}_*^{n\times p}$ called the *vertical space* at Y, and is denoted by $\mathcal{V}_Y = \left\{Y\Omega|\Omega^T = -\Omega, \Omega \in \mathbb{R}^{p\times p}\right\}$, see [96].

Define

$$\pi : \mathbb{R}_*^{n\times p} \to S_+^{n,p}$$
$$Y \mapsto YY^T.$$

Then π is invariant under the equivalence relation and induces a bijection γ on $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$ such that $\pi = \gamma \circ \tilde{\pi}$. For any function $\mathcal{E}(X)$ defined on $S_+^{n,p}$, there is a function F defined on $\mathbb{R}_*^{n\times p}$ that induces $\mathcal{E}$: for any $X = YY^T \in S_+^{n,p}$, $F(Y) := \mathcal{E} \circ \pi(Y) = \mathcal{E}(YY^T)$. This is summarized in the diagram below:

$$\begin{tikzcd}
\mathbb{R}_*^{n\times p} \arrow[d, "\tilde{\pi}"] \arrow[dr, dashed, "\pi=\gamma\circ\tilde{\pi}"] & & \\
\mathbb{R}_*^{n\times p}/\mathcal{O}_p \arrow[r, leftrightarrow, "\gamma"] & S_+^{n,p} \arrow[r, "\mathcal{E}"] & \mathbb{R}
\end{tikzcd}$$

In particular, $\mathcal{S}_+^{n,p}$ is diffeomorphic to $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$ under γ, see [96]. For any $Y \in \mathbb{R}_*^{n\times p}$, the flat metric for the total space $\mathbb{R}_*^{n\times p}$,

$$g(U,V) = \mathrm{Tr}(U^TV), \forall U,V \in T_Y\mathbb{R}_*^{n\times p} = \mathbb{R}^{n\times p}$$

induces a metric on the quotient manifold $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$ through Riemannian submersion, which is called Bures-Wasserstein metric, see [70, 69, 96].

The explicit expression of g_{BW} on $\mathcal{S}_+^{n,p}$ is as follows:

$$\begin{aligned} g_{BW}(U,V) = \mathrm{Tr}(uv^T) \qquad & \forall U,V \in T_XS_+^{n,p}, u,v \in T_Y\mathbb{R}_*^{n\times p} \\ & \text{s.t. } \mathrm{d}\pi(u) = U, \mathrm{d}\pi(v) = V, u,v \in \mathrm{Ker}(\mathrm{d}\pi)^\perp \end{aligned} \tag{3.3.14}$$

where X has decomposition $X = YY^T$, $\mathrm{d}\pi(u) = Yu^T + uY^T$ is the differential of π, and $u \in \mathrm{Ker}(\mathrm{d}\pi)^\perp \Leftrightarrow Y^Tu = u^TY$. The conditions on u, v are equivalent to that u, v are the *horizontal lift* of U, V respectively. Details will be discussed in proof of theorem 15.

For g_{BW}, we have the following theorem describing a $\mathbb{R}_*^{n\times p}$-valued process Y_t, such that $X_t = Y_tY_t^T$ gives the solution to SDE (3.1.2) in $(\mathcal{S}_+^{n,p}, g_{BW})$.

Theorem 15. *Let $Y_t = (Y_{ij})$ solves the following SDE in $\mathbb{R}_*^{n\times p}$*

$$\mathrm{d}Y_{ij} = -\frac{\partial\Phi(Y)}{\partial Y_{ij}}\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}W_t^{ij} \qquad 1 \le i \le n, 1 \le j \le p \tag{3.3.15}$$

where $\Phi(Y) = \mathcal{E}(YY^T) + \frac{1}{\beta}S(Y)$ is the energy function with a correction term, $S(Y) = \frac{1}{2}\sum_{i=1}^p\sum_{j=i+1}^p \log(\sigma_i^2 + \sigma_j^2)$, $\{\sigma_i\}_{i=1}^p$ are p singular values of Y.

Then $X_t = Y_tY_t^T$ solves SDE (3.1.2) in $(\mathcal{S}_+^{n,p}, g_{BW})$.

Lemma 13. *The gradient of the correction term S at the point $Y = Q\Sigma P^T$ is given* *by*

$$\mathrm{grad}S(Y) = Q \begin{bmatrix} \sum_{j\neq 1} \frac{\sigma_1}{\sigma_1^2+\sigma_j^2} & 0 & \cdots & 0 \\ 0 & \sum_{j\neq 2} \frac{\sigma_2}{\sigma_2^2+\sigma_j^2} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \sum_{j\neq p} \frac{\sigma_p}{\sigma_p^2+\sigma_j^2} \\ 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 \end{bmatrix} P^T \tag{3.3.16}$$

Proof. The differential of S is given by

$$\begin{aligned} \mathrm{d}S =& \frac{1}{2}\sum_{i=1}^{p}\sum_{j=i+1}^{p} \frac{2\sigma_i \mathrm{d}\sigma_i + 2\sigma_j \mathrm{d}\sigma_j}{\sigma_i^2+\sigma_j^2} \\ =& \sum_{i=1}^{n}(\sum_{j\neq i} \frac{\sigma_i}{\sigma_i^2+\sigma_j^2})\mathrm{d}\sigma_i \end{aligned} \tag{3.3.17}$$

Therefore, the derivative (also the gradient) of S in represented by matrix is given by

$$\mathrm{grad}S(Y) = Q \begin{bmatrix} \sum_{j\neq 1} \frac{\sigma_1}{\sigma_1^2+\sigma_j^2} & 0 & \cdots & 0 \\ 0 & \sum_{j\neq 2} \frac{\sigma_2}{\sigma_2^2+\sigma_j^2} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \sum_{j\neq p} \frac{\sigma_p}{\sigma_p^2+\sigma_j^2} \\ 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 \end{bmatrix} P^T \tag{3.3.18}$$

for $Y = Q\Sigma P^T$ $\square$

Remark 22. For g_E and g_{BW}, the volume form $\mathrm{d}V$ are different, so for any test function $f \in C^\infty(\mathcal{S}_+^{n,p})$, the expectation of f with respect to ρ_* may not be the same.

Proof of theorem 15. We first prove the expression of Bures-Wasserstein metric through Riemannian submersion. Recall that Riemannian submersion requires that $\mathrm{d}\pi_Y$: $\mathrm{Ker}(\mathrm{d}\pi_Y)^\perp \to T_{\pi(Y)}\mathcal{S}_+^{n,p}$ gives an isometry for each $Y \in \mathbb{R}_*^{n\times p}$, so understanding of $\mathrm{Ker}(\mathrm{d}\pi)^\perp$ is essential.

Since $\pi(Y) = YY^T$, $\mathrm{d}\pi_Y(u) = Yu^T + uY^T$, $\mathrm{Ker}(\mathrm{d}\pi_Y) = \{v \in \mathbb{R}^{n\times p} | Yv^T + vY^T = 0\}$. $\mathrm{Ker}(\mathrm{d}\pi)^\perp$ is therefore defined by

$$\mathrm{Ker}^\perp(\mathrm{d}\pi_Y) = \left\{u \in \mathbb{R}^{n\times p} \middle| \mathrm{Tr}(uv^T) = 0, Yv^T + vY^T = 0\right\} \tag{3.3.19}$$

For each $u \in \mathrm{Ker}^\perp(\mathrm{d}\pi_Y)$, take $v = Y(Y^Tu - u^TY) \in \mathrm{Ker}(\mathrm{d}\pi_Y)$, we have

$$\mathrm{Tr}(uv^T) = \frac{1}{2}\mathrm{Tr}(uv^T + vu^T) = -\frac{1}{2}\mathrm{Tr}\big((Y^Tu - u^TY)(Y^Tu - u^TY)^T\big) = 0 \tag{3.3.20}$$

so

$$\mathrm{Ker}^\perp(\mathrm{d}\pi_Y) = \left\{u \in \mathbb{R}^{n\times p} \middle| Y^Tu = u^TY\right\} \tag{3.3.21}$$

The proof of theorem 15 is the direct application of 13. When Y_t is a Brownian motion on $\mathbb{R}^{n\times p}$, $X_t = \pi(Y_t)$ solves RLE on $(\mathcal{S}_+^{n,p}, g_{BW})$ with $\mathcal{E} = -\frac{1}{\beta}S$. We also need the fact that, for any $\mathcal{E} \in C^\infty(\mathcal{M})$, where $\pi : (\widetilde{\mathcal{M}}, \tilde{g}) \to (\mathcal{M}, g)$ is a Riemannian submersion, $\mathrm{d}\pi(\tilde{\mathrm{grad}}\,\tilde{\mathcal{E}}) = \mathrm{grad}\,\mathcal{E}$, i.e. there is no different between taking gradient of a function on $\mathcal{M}$ or $\widetilde{\mathcal{M}}$.

Thus, if Y_t solves RLE of $\mathcal{E}' = \tilde{\mathcal{E}} + \frac{1}{\beta}\tilde{S}$, $\pi(Y_t)$ solves RLE on $(\mathcal{S}_+^{n,p}, g_{BW})$ of energy

$\mathcal{E}$. The fact that we can just add the Brownian motion term and gradient flow term to get a RLE should be understood in the generator sense, i.e. the generator of RLE is the sum of generators of a Brownian motion and a gradient flow.

What is left now is to check the condition for mean curvature $\boldsymbol{H}$, and compute the function $S(X)$. In the rest of the proof, we are not using the assertion of [78] that $S(X) = \log \mathrm{vol}(\pi^{-1}(X))$ but analyzing $\boldsymbol{H}$ from its definition. To do this we will construct an orthonormal basis $\{\boldsymbol{w}_I\}_{I\in\mathcal{I}} \cup \{\boldsymbol{v}_J\}_{J\in\mathcal{J}} \subset \mathfrak{X}(T\mathbb{R}_*^{n\times p})$, such that $\mathrm{Span}(\{\boldsymbol{v}_J\}_{J\in\mathcal{J}}) = \mathrm{Ker}(\mathrm{d}\pi)$ and $\mathrm{Span}(\{\boldsymbol{w}_I\}_{I\in\mathcal{I}}) = \mathrm{Ker}(\mathrm{d}\pi)^{\perp}$. By dimension counting we have $|\mathcal{J}| = \dim \mathcal{O}_p = \frac{p(p-1)}{2}$, $|\mathcal{I}| = np - \frac{p(p-1)}{2}$.

For $Y \in \mathbb{R}_*^{n\times p}$, its singular value decomposition $Y = Q\Sigma P^T$ always exists, where $Q \in \mathcal{O}_n, P \in \mathcal{O}_p$ and $\Sigma = \mathrm{diag}(\sigma_1, ..., \sigma_p, 0, ..., 0)$, $\sigma_1 \geq \cdots \geq \sigma_p > 0$. Define the following vector fields $\{\boldsymbol{u}_{ij}\}_{1\leq i\leq p, i+1\leq j\leq n}$, $\{\boldsymbol{s}_i\}_{i=1}^p$ and $\{\boldsymbol{v}_{ij}\}_{1\leq i<j\leq p}$, and we use their action on Y matrix to describe them, i.e.

$$
\begin{aligned}
u_{ij} &:= \boldsymbol{u}_{ij}(Y) = Q\Big(\frac{\sigma_j}{\sqrt{\sigma_i^2+\sigma_j^2}} e_i\varepsilon_j^T + \frac{\sigma_i}{\sqrt{\sigma_i^2+\sigma_j^2}} e_j\varepsilon_i^T\Big)P^T && 1 \leq i < j \leq p \\
u_{ij} &:= \boldsymbol{u}_{ij}(Y) = Q e_j \varepsilon_i^T P^T && 1 \leq i \leq p, p+1 \leq j \leq n \\
s_i &:= \boldsymbol{s}_i(Y) = Q e_i \varepsilon_i^T P^T && 1 \leq i \leq p \\
v_{ij} &:= \boldsymbol{v}_{ij}(Y) = Q\Big(\frac{\sigma_i}{\sqrt{\sigma_i^2+\sigma_j^2}} e_i\varepsilon_j^T - \frac{\sigma_j}{\sqrt{\sigma_i^2+\sigma_j^2}} e_j\varepsilon_i^T\Big)P^T && 1 \leq i < j \leq p
\end{aligned}
\tag{3.3.22}
$$

where e_i is $n \times 1$, i-th standard unit vector, and ε_i is $p \times 1$, i-th standard unit vector.

By checking that this basis is orthonormal and that $\boldsymbol{v}_{ij}(YY^T) = \boldsymbol{v}_{ij}(Y)Y^T + Y\boldsymbol{v}_{ij}(Y^T) = 0$, we can see that $\{\boldsymbol{w}_I\}_{I\in\mathcal{I}} = \{\boldsymbol{u}_{ij}\}_{1\leq i\leq p, i+1\leq j\leq n} \cup \{\boldsymbol{s}_i\}_{i=1}^p$ and $\{\boldsymbol{v}_J\}_{J\in\mathcal{J}} = \{\boldsymbol{v}_{ij}\}_{1\leq i<j\leq p}$ are the desired vector fields.

Now we need to compute the mean curvature $\boldsymbol{H}(Y)$ of $\mathcal{O}_Y = \pi^{-1}(\pi(Y))$ and check that $\mathrm{d}\pi(\boldsymbol{H}(Y))$ depends only on $\pi(Y)$. By definition,

$$\boldsymbol{H}(Y) = \sum_{J\in\mathcal{J}} (\mathcal{D}_{\boldsymbol{v}_J}\boldsymbol{v}_J)^{\perp}(Y) \tag{3.3.23}$$

where $(\cdot)^{\perp}$ is the component in $N\mathcal{O}_Y = (T\mathcal{O}_Y)^{\perp}$.

What remains is to compute $\mathcal{D}_{\boldsymbol{v}_J}\boldsymbol{v}_J$, by either Koszul's formula or direct computing derivative, and we will use the latter. As all Christoffel symbols vanishes in Euclidean space under entry-wise basis, the matrix $\mathcal{D}_{\boldsymbol{v}_J}\boldsymbol{v}_J(Y) = \boldsymbol{v}_J(\boldsymbol{v}_J(Y))$. Therefore the action of $\boldsymbol{v}_J$ on Q, Σ, P^T of the SVD of Y is needed. In fact, we have

$$\boldsymbol{v}_{ij}(Q) = 0,\ \boldsymbol{v}_{ij}(\Sigma) = 0,\ \boldsymbol{v}_{ij}(P^T) = \frac{1}{\sqrt{\sigma_i^2+\sigma_j^2}}(\varepsilon_i\varepsilon_j^T - \varepsilon_j\varepsilon_i^T)P^T \qquad 1 \le i < j \le p \tag{3.3.24}$$

so

$$\begin{aligned}(\mathcal{D}_{\boldsymbol{v}_{ij}}\boldsymbol{v}_{ij})(Y) = \boldsymbol{v}_{ij}(\boldsymbol{v}_{ij}(Y)) =& \boldsymbol{v}_{ij}(Q(\frac{\sigma_i}{\sqrt{\sigma_i^2+\sigma_j^2}}e_i\varepsilon_j^T - \frac{\sigma_j}{\sqrt{\sigma_i^2+\sigma_j^2}}e_j\varepsilon_i^T)P^T) \\ =& Q(\frac{\sigma_i}{\sqrt{\sigma_i^2+\sigma_j^2}}e_i\varepsilon_j^T - \frac{\sigma_j}{\sqrt{\sigma_i^2+\sigma_j^2}}e_j\varepsilon_i^T)\boldsymbol{v}_{ij}(P^T) \\ =& Q(\frac{\sigma_i}{\sqrt{\sigma_i^2+\sigma_j^2}}e_i\varepsilon_j^T - \frac{\sigma_j}{\sqrt{\sigma_i^2+\sigma_j^2}}e_j\varepsilon_i^T)\frac{1}{\sqrt{\sigma_i^2+\sigma_j^2}}(\varepsilon_i\varepsilon_j^T - \varepsilon_j\varepsilon_i^T)P^T \\ =& -\frac{1}{\sigma_i^2+\sigma_j^2}Q(\sigma_i e_i\varepsilon_i + \sigma_j e_j\varepsilon_j)P^T = -\frac{\sigma_i}{\sigma_i^2+\sigma_j^2}s_i - \frac{\sigma_j}{\sigma_i^2+\sigma_j^2}s_j\end{aligned} \tag{3.3.25}$$

an important observation here is that $\boldsymbol{s}_i$ is the gradient of the function $\sigma_i(Y)$, i.e. $\boldsymbol{s}_i = \mathcal{D}(\sigma_i)$, therefore by chain rule

$$\mathcal{D}_{\boldsymbol{v}_{ij}}\boldsymbol{v}_{ij} = -\mathcal{D}(\frac{1}{2}\log(\sigma_i^2+\sigma_j^2)) \tag{3.3.26}$$

As σ_i is a function defined on $\mathcal{S}_+^{n,p}$, for $X = YY^T$, $Y = Q\Sigma P^T$, define $\tilde{S}(Y) = S(X) = \sum_{1\leq i<j\leq p} \frac{1}{2}\log(\sigma_i^2 + \sigma_j^2)$, we have

$$\boldsymbol{H} = -\mathcal{D}\tilde{S}\,, \qquad \boldsymbol{h} = \mathrm{d}\pi(\boldsymbol{H}) = -\mathrm{grad}_{g_{BW}} S \tag{3.3.27}$$

$\square$

3.4 Convergence analysis of RLE

3.4.1 Convergence to Gibbs

We first provide some Riemannian version of standard results on Langevin equation (3.1.2): the Fokker-Planck equation which describes the evolution of law ρ_t of $\boldsymbol{X}_t$, and the convergence of ρ_t to Gibbs distribution.

Proposition 2. *Let ρ_t be the law of $\boldsymbol{X}_t$, where $\boldsymbol{X}_t$ solves RLE* (3.1.2)*, then the evolution of ρ_t satisfies the following Fokker-Planck equation*

$$\partial_t \rho_t = \mathrm{div}(\frac{1}{\beta}\mathrm{grad}\rho_t + \rho_t\, \mathrm{grad}\mathcal{E}) = \frac{1}{\sqrt{g}}\partial_i(\sqrt{g} g^{ij}(\frac{1}{\beta}\partial_j\rho_t + \rho_t\partial_j\mathcal{E})) \tag{3.4.1}$$

where g^{ij} is the inverse of metric g_{ij} and $g = \det(g_{ij})$.

The convergence of Fokker-Planck type equation has been studied extensively in literature [68, 3, 89], and a general convergence result based on entropy dissipation is the following:

Proposition 3. *Suppose the energy function $\mathcal{E}$ satisfies that*

$$Z_\beta := \int_{\mathcal{M}} e^{-\beta\mathcal{E}} \mathrm{d}V < \infty \tag{3.4.2}$$

so $\rho_* = \frac{1}{Z_\beta} e^{-\beta\mathcal{E}}$ defines a probability measure on $\mathcal{M}$.

Take any function $\phi \in C^2(\mathbb{R}_+)$ such that $\phi(\cdot) \geq \phi(1) = 0, \phi''(\cdot) > 0$, then $\mathcal{F}(\rho) := \int_{\mathcal{M}} \phi(\frac{\rho}{\rho_*})\rho_* \mathrm{d}V$, where $\mathrm{d}V = \sqrt{g}\mathrm{d}x_1 \cdots \mathrm{d}x_n$ is the volume form, defines a

Lyapunov functional for (3.4.1)*, and we have that*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(\rho_t) =& \frac{1}{\beta}\int_{\mathcal{M}}\phi'(\frac{\rho_t}{\rho_*})\mathrm{div}(\rho_*\mathrm{grad}\,\frac{\rho_t}{\rho_*})\mathrm{d}V \\ =& -\frac{1}{\beta}\int_{\mathcal{M}}\rho_*\phi''(\frac{\rho}{\rho_*})|\mathrm{grad}\,\frac{\rho}{\rho_*}|_g^2\mathrm{d}V < 0\end{aligned} \tag{3.4.3}$$

where $|\mathrm{grad}\, f|_g^2 = g^{ij}\partial_i f\partial_j f$ *is the norm of* $\mathrm{grad}\, f$.

Thus, the convergence $\lim_{t\to\infty}\mathcal{F}(\rho_t) = 0$ *is assured.*

The candidates of ϕ includes a wide class of functions with $\phi(h) = h\log h - h + 1$ and $\phi(h) = (h-1)^2$ being two extremals [68]. For the first choice $\phi(h) = h\log h - h + 1$ we get the convergence of KL-divergence

$$\lim_{t\to\infty}\mathrm{D}_{KL}(\rho_t||\rho_*) = 0 \tag{3.4.4}$$

so the total variation distance $\delta(\rho_t, \rho_*)$ also converges to 0 by Pinsker's inequality $\delta(P,Q) \leq \sqrt{\frac{1}{2}\mathrm{D}_{KL}(P,Q)}$. For the other choice $\phi(h) = (h-1)^2$ we get the L^2 convergence of $\frac{\rho_t}{\sqrt{\rho_*}}$ to $\sqrt{\rho_*}$.

$$\lim_{t\to\infty}\int_{\mathcal{M}}(\frac{\rho_t}{\rho_*} - 1)^2\rho_*\mathrm{d}V = \lim_{t\to\infty}\int_{\mathcal{M}}(\frac{\rho_t}{\sqrt{\rho_*}} - \sqrt{\rho_*})^2\mathrm{d}V = 0 \tag{3.4.5}$$

3.4.2 Rate convergence: review of general results

Now as the convergence is guaranteed, we are more concerned about rate convergence. A detailed analysis on the condition for rate convergence is beyond the scope of this paper. In this section we just review some classical results on the convergence of Fokker-Planck type equations and provide some simple modification on Riemannian manifolds.

It is well known that the following logarithmic Sobolev inequality implies dissipation of KL-divergence ($\phi(h) = h\log h - h + 1$) in a exponential rate

$$\int_{\mathcal{M}} f^2 \log f^2 \mathrm{d}\mu - \int_{\mathcal{M}} f^2 \mathrm{d}\mu (\log \int_{\mathcal{M}} f^2 \mathrm{d}\mu) \leq \frac{2}{\lambda} \int_{\mathcal{M}} |\mathrm{grad} f|_g^2 \mathrm{d}\mu \,, \tag{3.4.6}$$

where $\mathrm{d}\mu = \rho_* \mathrm{d}V$ is a probability measure on $\mathcal{M}$, $f \in C^\infty(\mathcal{M})$, $\lambda > 0$ is a constant. Then take $f = \sqrt{\frac{\rho_t}{\rho_*}}$ we have

$$\mathrm{D}_{KL}(\rho_t || \rho_*) \leq \frac{2}{\lambda} \int_{\mathcal{M}} \rho_* |\mathrm{grad} \sqrt{\frac{\rho_t}{\rho_*}}|_g^2 \mathrm{d}V = -\frac{\beta}{2\lambda} \frac{\mathrm{d}}{\mathrm{d}t} \mathrm{D}_{KL}(\rho_t || \rho_*) \tag{3.4.7}$$

which leads to $\mathrm{D}_{KL}(\rho_t || \rho_*) \leq C \mathrm{e}^{-\frac{2\lambda}{\beta} t}$

On a Riemannian manifold it is well known that the logarithmic Sobolev inequality is implied by the Bakry-Émery criteria [9], which can be stated as in [68]:

Theorem 16. *Let $e^{-\mathcal{E}}$ be a probability measure on a Riemannian manifold $\mathcal{M}$, Hess$\,\mathcal{E}$ + Ric $\geq \lambda g$, where g is the metric, Hess is the Riemannian Hessian, and Ric is the Ricci curvature tensor on the manifold. Then, $e^{-\mathcal{E}}$ satisfies a logarithmic Sobolev inequality with constant λ.*

In the literature, $\mathrm{Ric}_\infty = \mathrm{Ric} + \mathrm{Hess}\,\mathcal{E}$ is also called the Bakry-Émery-Ricci tensor, and a triple $(\mathcal{M}, g, \mathcal{E})$ is often called Bakry-Émery manifold, where the measure on $\mathcal{M}$ is $\mathrm{e}^{(-\mathcal{E})} \mathrm{d}V$.

For the Gibbs distribution $\mathrm{e}^{-\beta\mathcal{E}}$, the Bakry-Émery-Ricci tensor should be refined accordingly. The following calculation was given in [62]:

Lemma 14. *For the Gibbs measure $e^{-\beta\mathcal{E}}$, the second order carré du champ operator is*

$$\Gamma_2(f,f) = \frac{1}{\beta^2} |Hess\, f|_g^2 + \frac{1}{\beta^2} Ric_g(\mathrm{grad} f, \mathrm{grad} f) + \frac{1}{\beta} Hess\, \mathcal{E}(\mathrm{grad} f, \mathrm{grad} f).$$

Thus the Bakry-Émery-Ricci tensor should be written as [62]:

$$\mathrm{Ric}_\infty = \frac{1}{\beta}\mathrm{Ric} + \mathrm{Hess}\,\mathcal{E}.$$

And Bakry-Émery's curvature dimension condition is

$$\frac{1}{\beta}\mathrm{Ric}(h,h) + \mathrm{Hess}\,\mathcal{E}(h,h) \geq \rho g(h,h), \quad \forall h \in T_X\mathcal{S}_+^{n,p},$$

where g is the metric, and $\rho > 0$ can be a function.

If $\rho \geq \lambda > 0$ for a constant λ, then we obtain logarithmic Sobolev inequality with contant λ, which further implies the exponential convergence to equilibrium with speed λ.

In general, to obtain a sharp estimate of lower bound of ρ, it is quite difficult. For instance, even the Riemannian Hessian is nontrival to estimate [96] for an arbitary energy function $\mathcal{E}$.

Remark 23. Hopf–Rinow theorem states that the geodesic completeness is equivalent to the metric space completeness for a connected smooth Riemannian manifold. So $\mathcal{M}$ is not complete thus not compact, e.g., a sequence of rank-2 matrices can converge to a rank-1 matrix. Many results, especially those from differential geometry and Riemannian geometry, are often for a compact manifold. If needed, we may consider a compact modification of $\mathcal{S}_+^{n,p}$. For example, let $\mathcal{M} = \{X \in \mathbb{R}^{n\times n} | X = X^T, X \succeq 0, \mathrm{rank}(X) = p\}$ and we can also consider a compact version of it $\mathcal{M}_\varepsilon = \{X \in \mathbb{R}^{n\times n} | X = X^T, X \succeq 0, \mathrm{rank}(X) = p, \quad 0 \leq \lambda_p \leq \lambda_{p-1} \leq \cdots \leq \lambda_1 \leq \frac{1}{\varepsilon}\}$ where $\varepsilon > 0$ is a very small number. It looks possible to modify and extend the results in this paper to such a compact manifold. But for convenience, we only discuss the original manifold $\mathcal{S}_+^{n,p}$ without considering any compactness.

The Bakry-Émery criteria for $\mathcal{S}_+^{n,p}$ under two metrics will be discussed in the following sections. Below we discuss some cases beyond Bakry-Émery condition.

1. The standard Holley-Stroock perturbation Lemma [42] applies in our case. A simple proof can be found in [59, Lemma 1.2], with a probability measure on $\mathbb{R}^n$ replaced by a probability measure on $\mathcal{M}$.

   Lemma 15. *If a probability measure* $\mathrm{d}m = e^{-\beta\mathcal{E}}\mathrm{d}V \in \mathcal{P}(\mathcal{M})$ *satisfies a logarithmic Sobolev inequality with* λ*,*

   $$\int_{\mathcal{M}} f^2 \log f^2 \mathrm{d}m - \int_{\mathcal{M}} f^2 \mathrm{d}m \log(\int_{\mathcal{M}} f^2 \mathrm{d}m) \le \frac{1}{\lambda}\int_{\mathcal{M}} |\mathrm{grad}\, f|_g^2 \mathrm{d}m \tag{3.4.8}$$

   then for another probability measure $\mathrm{d}m' = e^{-\beta(\mathcal{E}+e)}$ *with* $e \in L^\infty(\mathcal{M})$*, it satisfies a logarithmic Sobolev inequality with* $\lambda e^{-2\beta\,\mathrm{osc}\, e}$*, where* $\mathrm{osc}\, e = \mathrm{sup}_{\mathcal{M}} e - \mathrm{inf}_{\mathcal{M}} e$*.*

2. In [89, Theorem 3], a polynomial decay of the form $\mathrm{D}_{KL}(\rho_t||\rho_*) \le \frac{C}{t^\kappa}$ for any $\kappa > 0$ is built with growth conditions on the energy function $\mathcal{E}$. The key construction is a bound $H(f|\rho_*) \le CI(f|\rho_*)^{1-\delta}M_s(f)^\delta$, where $s > 0$ is arbitrary, $\delta > 0$ depends on s

   $$\begin{aligned} H(f|\rho_*) &= \int f \log\frac{f}{\rho_*}, \quad I(f|\rho_*) = \int f|\frac{\nabla f}{f} - \frac{\nabla \rho_*}{\rho_*}|^2 \\ M_s(f) &= \int f(x)(1+|x|^2)^{\frac{s}{2}}\mathrm{d}x\,, \end{aligned} \tag{3.4.9}$$

   and a slow growth condition on $M_s(\rho_t)$

   $$M_s(\rho_t) \le M_s(\rho_0) + C_s t \tag{3.4.10}$$

   This theorem is for Euclidean case only, but a Riemannian version is also expected.

3. The Fokker-Planck equation (3.4.1) ($\beta = 1$)can be revised to

$$\partial_t p_t = \nabla \cdot (\boldsymbol{D}(\nabla p_t + p_t \nabla E)) \tag{3.4.11}$$

with $\boldsymbol{D} = (g^{ij})$ being the inverse metric under a coordinate chart $x : \mathcal{M} \to \Omega$, and $p_t = \sqrt{g}\rho_t$, $E = \mathcal{E} - \log\sqrt{g}$. In [3, section 2], the exponential L^2 convergence of $\frac{p_t}{\sqrt{p_*}}$ to $\sqrt{p_*}$ for the equation above was studied and the condition is the spectrum gap of a Hamiltonian $\mathcal{H}$

$$\mathcal{H}(f, f) = \int_\Omega \nabla(\frac{f}{\sqrt{p_*}})^T \boldsymbol{D} \nabla(\frac{f}{\sqrt{p_*}}) p_* \mathrm{d}x \tag{3.4.12}$$

As $\sqrt{p_*}$ is apparently the ground state with energy 0, if the spectrum $\sigma(\mathcal{H})$ of $\mathcal{H}$ has a positive spectrum gap, i.e. the positive part of spectrum $\sigma(\mathcal{H})$ has a positive distance λ_0 to 0, then we have exponential decay of $||\frac{p_t}{\sqrt{p_*}} - \sqrt{p_*}||_{L^2(\Omega)}$ with rate $\mathrm{e}^{-\lambda_0 t}$.

3.4.3 Bakry-Émery-Ricci tensor for the embedded geometry

In this section, we will see that the Ricci tensor of $(\mathcal{S}_+^{n,p}, g_E)$ is indefinite in the region where some of the eigenvalues are much smaller than others, which in fact breaks the Bakry-Émery's condition.

In the evaluation of Ric_∞ for embedded manifold $\mathcal{S}_+^{n,p}$, the second fundamental form $\boldsymbol{I\!I}$ plays a central role and explanations are as follows.

Using flatness of the ambient space $\mathrm{Sym}(n)$, Riemannian curvature tensor could be expressed by Gauss-Codazzi equation using only second fundamental form, so Ricci curvature tensor is known. Also, the Riemannian Hessian of any energy function $\mathcal{E} \in C^\infty(\mathcal{S}_+^{n,p})$ which is the restriction of a function $\tilde{\mathcal{E}} \in C^\infty(\mathrm{Sym})$ defined in the ambient space, can be expressed in terms of second fundamental form $\boldsymbol{I\!I}$ along with

Hessian and gradient of $\tilde{\mathcal{E}}$. As follows are the two Lemmas:

Lemma 16. *[Gauss-Codazzi Equation] Let* $(\mathcal{M}, g)$ *be a embedded manifold of* $(\widetilde{\mathcal{M}}, \tilde{g})$, *i.e.* $\mathcal{M} \subset \widetilde{\mathcal{M}}$ *and* $g(\boldsymbol{v}, \boldsymbol{v}) = \tilde{g}(\boldsymbol{v}, \boldsymbol{v})$ *for all* $\boldsymbol{v} \in \mathfrak{X}(\mathcal{M})$. *Let* $R, \tilde{R}$ *be the Riemannian curvature tensor of* $\mathcal{M}$ *and* $\widetilde{\mathcal{M}}$ *respectively,* $\boldsymbol{I\!I} : \mathfrak{X}(\mathcal{M}) \otimes \mathfrak{X}(\mathcal{M}) \to \mathfrak{X}(\mathcal{M})^{\perp}$ *be the second fundamental form of* $\mathcal{M}$, *then for* $X, Y, Z, W \in \mathfrak{X}(\mathcal{M})$

$$\langle \tilde{R}(X,Y)Z, W \rangle = \langle R(X,Y)Z, W \rangle - \langle \boldsymbol{I\!I}(X,W), \boldsymbol{I\!I}(Y,Z) \rangle + \langle \boldsymbol{I\!I}(X,Z), \boldsymbol{I\!I}(Y,W) \rangle \,, \tag{3.4.13}$$

where $\langle \cdot, \cdot \rangle = \tilde{g}(\cdot, \cdot)$.

Lemma 17. *Let* $(\mathcal{M}, g)$ *be a embedded manifold of* $(\widetilde{\mathcal{M}}, \tilde{g})$, *with a function* $\tilde{f} \in C^{\infty}(\widetilde{\mathcal{M}})$ *whose restriction on* $\mathcal{M}$ *is denoted by* f. *For* $\boldsymbol{v} \in \mathfrak{X}(\mathcal{M})$, *the Riemannian Hessian* $\tilde{\operatorname{Hess}}\, \tilde{f}(\boldsymbol{v}, \boldsymbol{v})$ *and* $\operatorname{Hess} f(\boldsymbol{v}, \boldsymbol{v})$ *satisfies*

$$\operatorname{Hess} f(\boldsymbol{v}, \boldsymbol{v}) = \tilde{\operatorname{Hess}}\, \tilde{f}(\boldsymbol{v}, \boldsymbol{v}) + \langle \boldsymbol{I\!I}(\boldsymbol{v}, \boldsymbol{v}), \operatorname{grad} \tilde{f} \rangle \,. \tag{3.4.14}$$

Proof. Let $\gamma : [-1, 1] \to \mathcal{M}$ be a smooth curve such that $\dot{\gamma}(t) = \boldsymbol{v}(\gamma(t))$ holds for $t \in (-\epsilon, \epsilon)$ with some $\epsilon > 0$. Denoted by $\tilde{\gamma} : [-1, 1] \to \widetilde{\mathcal{M}}, \tilde{\gamma}(t) = \gamma(t)$ the same curve in the ambient space, then $f(\gamma(t)) = \tilde{f}(\tilde{\gamma}(t))$. Taking second order derivative on both sides we get

$$\ddot{f}(\gamma(t)) = \dot{(}\langle \boldsymbol{v}, \operatorname{grad} f \rangle) = \langle \nabla_{\boldsymbol{v}} \boldsymbol{v}, \operatorname{grad} f \rangle + \operatorname{Hess} f(\boldsymbol{v}, \boldsymbol{v}) \tag{3.4.15}$$

$$= \ddot{\tilde{f}}(\tilde{\gamma}(t)) = \dot{(}\langle \boldsymbol{v}, \tilde{\operatorname{grad}}\, \tilde{f} \rangle) = \langle \tilde{\nabla}_{\boldsymbol{v}} \boldsymbol{v}, \tilde{\operatorname{grad}}\, \tilde{f} \rangle + \tilde{\operatorname{Hess}}\, \tilde{f}(\boldsymbol{v}, \boldsymbol{v}) \,. \tag{3.4.16}$$

As $\boldsymbol{I\!I}(\boldsymbol{v}, \boldsymbol{v}) = (\tilde{\nabla}_{\boldsymbol{v}} \boldsymbol{v})^{\perp} = \tilde{\nabla}_{\boldsymbol{v}} \boldsymbol{v} - \nabla_{\boldsymbol{v}} \boldsymbol{v}$, and for any $\boldsymbol{u} \in \mathfrak{X}(\mathcal{M})$, $\langle \boldsymbol{u}, \operatorname{grad} f \rangle =$

$\langle \boldsymbol{u}, \tilde{\operatorname{grad}} \tilde{f} \rangle$, we have

$$
\begin{aligned}
\langle \widetilde{\nabla}_{\boldsymbol{v}} \boldsymbol{v}, \tilde{\operatorname{grad}} \tilde{f} \rangle =& \langle \nabla_{\boldsymbol{v}} \boldsymbol{v} + \boldsymbol{I\!I}(\boldsymbol{v}, \boldsymbol{v}), \tilde{\operatorname{grad}} \tilde{f} \rangle \\
=& \langle \nabla_{\boldsymbol{v}} \boldsymbol{v}, \operatorname{grad} f \rangle + \langle \boldsymbol{I\!I}(\boldsymbol{v}, \boldsymbol{v}), \operatorname{grad} \tilde{f} \rangle \,. \qquad (3.4.17)
\end{aligned}
$$

□

Corollary 3. *If* $\tilde{\operatorname{grad}} \tilde{f} = \operatorname{grad} f$*, then* $\operatorname{Hess} f$ *is just the restriction of* $\tilde{\operatorname{Hess}} \tilde{f}$*. In particular, for* $f = \frac{1}{2}||X||_F^2 = \sum_{i=1}^p \lambda_i^2 \in C^\infty(\mathcal{S}_+^{n,p})$*, we have* $\operatorname{Hess} f = g_E$*. (i.e.* $\operatorname{Hess} f(\boldsymbol{v}, \boldsymbol{v}) = g_E(\boldsymbol{v}, \boldsymbol{v})$*).*

Remark 24. [96, Proposition 3.7] has given the expression of Riemannian Hessian $\operatorname{Hess} f$ for the embedded geometry:

$$
\operatorname{Hess} f(X)[\xi_X] = P_X^t(\nabla^2 f(X)[\xi_X]) + P_X^p \left(\nabla f(X)(X^\dagger \xi_X^p)^* + (\xi_X^p X^\dagger)^* \nabla f(X) \right) .
$$

In particular, for $f(X) = \frac{1}{2} \|X\|_F^2$, by plugging f in the expression one can verify that $\operatorname{Hess} f(X)$ is indeed the identity mapping.

Corollary 4. *At* $X \in \mathcal{S}_+^{n,p}$*, with* $X = Q \Lambda Q^T$*. For any vector field* $\boldsymbol{v} \in \mathfrak{X}(T\mathcal{S}_+^{n,p})$*, the second fundamental form is given by*

$$
\boldsymbol{I\!I}_X(\boldsymbol{v}, \boldsymbol{v}) = P_X^\perp \big(Q[\omega, [\omega, \Lambda]] Q^T \big) \,, \qquad (3.4.18)
$$

where $\omega = Q^{-1} \boldsymbol{v}(\mathrm{d}Q)$ *is an anti-symmetric matrix represents the* Q*-component of* $\boldsymbol{v}$*,* $[A, B] = AB - BA$ *is the matrix commutator, and* $P_{\boldsymbol{X}}^\perp$ *is the projection operator to*

$(T_X\mathcal{S}_+^{n,p})^\perp$, *given by*

$$P_X^\perp(Q\begin{bmatrix} A_{p\times p} & B_{p\times(n-p)} \\ (B_{p\times(n-p)})^T & C_{(n-p)\times(n-p)} \end{bmatrix}Q^T) = Q\begin{bmatrix} 0_{p\times p} & 0_{p\times(n-p)} \\ (0_{p\times(n-p)})^T & C_{(n-p)\times(n-p)} \end{bmatrix}Q^T\,. \tag{3.4.19}$$

Proof. For any constant matrix $\eta \in (T_{X_0}\mathcal{S}_+^{n,p})$, $X_0 = Q_0\Lambda_0Q_0^T$, we prove that

$$\langle \boldsymbol{I\!I}_{X_0}(\boldsymbol{v},\boldsymbol{v}),\eta\rangle = \langle Q_0[\omega_0,[\omega_0,\Lambda_0]]Q_0^T,\eta\rangle\,. \tag{3.4.20}$$

To do this, consider an auxiliary function $\tilde{f}_\eta(\tilde{X}) = \mathrm{Tr}(\tilde{X}\eta) = \langle X,\eta\rangle \in C^\infty(\mathrm{Sym}(n))$. Under Euclidean metric $\tilde{g}$, the Hessian of $\tilde{f}$ vanishes, but the restiction f of $\tilde{f}$ may has non-vanishing Hessian.

Let $\gamma(t),\tilde{\gamma}(t)$ be the same as in the proof of Lemma 17, with $\gamma(0) = X_0$. Consider the secondary derivative of $\tilde{f}(\tilde{\gamma}(t))$

$$\ddot{\tilde{f}}(\tilde{\gamma}(t)) = \dot{(\langle \boldsymbol{v},\tilde{\mathrm{grad}}\,\tilde{f}\rangle)} = \langle \widetilde{\nabla}_{\boldsymbol{v}}\boldsymbol{v},\tilde{\mathrm{grad}}\,\tilde{f}\rangle + \tilde{\mathrm{Hess}}\,\tilde{f}(\boldsymbol{v},\boldsymbol{v})\,. \tag{3.4.21}$$

Notice that $\tilde{\mathrm{grad}}\,\tilde{f} = \eta$, and $\tilde{\mathrm{Hess}}\,\tilde{f} = 0$. When $t = 0$, $\eta \in (T_{\gamma(0)}\mathcal{S}_+^{n,p})^\perp$, so we find

$$\ddot{\tilde{f}}(\tilde{\gamma}(0)) = \langle \widetilde{\nabla}_{\boldsymbol{v}}\boldsymbol{v},\eta\rangle = \langle \boldsymbol{I\!I}_{X_0}(\boldsymbol{v},\boldsymbol{v}),\eta\rangle\,. \tag{3.4.22}$$

On the other hand, consider the secondary derivative of $f(\gamma(t)) = \mathrm{Tr}(Q(t)\Lambda(t)Q^T(t)\eta)$, with convention $\omega(t) = Q^{-1}(t)\dot{Q}(t)$, we have

$$\ddot{f}(\gamma(t)) = \dot{\Big(\mathrm{Tr}\big(Q(\dot{\Lambda}+[\omega,\Lambda])Q^T\eta\big)\Big)} = \mathrm{Tr}\big(Q(\ddot{\Lambda}+[\dot{\omega},\Lambda]+2[\omega,\dot{\Lambda}]+[\omega,[\omega,\Lambda]])Q^T\eta\big)\,. \tag{3.4.23}$$

Take $t = 0$, all terms in the parentheses except $Q[\omega, [\omega, \Lambda]]Q^T$ vanish as they are in $T_{X_0}\mathcal{S}_+^{n,p}$ while $\eta \in (T_{X_0}\mathcal{S}_+^{n,p})^\perp$, thus

$$\langle Q_0[\omega_0, [\omega_0, \Lambda_0]]Q_0^T, \eta\rangle = \ddot{f}(\gamma(0)) = \ddot{\tilde{f}}(\tilde{\gamma}(0)) = \langle \boldsymbol{I\!I}_{X_0}(\boldsymbol{v}, \boldsymbol{v}), \eta\rangle \tag{3.4.24}$$

□

Now we are able to compute Ricci tensor Ric for $(\mathcal{S}_+^{n,p}, g_E)$.

Theorem 17 (Ricci tensor for $\mathcal{S}_+^{n,p}$)**.** *Under the orthonormal basis $\{\boldsymbol{v}_I\}_{I\in\mathcal{I}}$, the Ricci tensor Ric of $(\mathcal{S}_+^{n,p}, g_E)$ is given by*

$$Ric(\boldsymbol{v}, \boldsymbol{v}) = \sum_{i=1}^{p} r_i(\sum_{j=p+1}^{n} c_{ij}^2) \tag{3.4.25}$$

for an arbitrary vector $\boldsymbol{v} = \sum_{i=1}^p c_i\boldsymbol{v}_i + \sum_{i=1}^p\sum_{j=i+1}^n c_{ij}\boldsymbol{v}_{ij}$, where $r_i = \frac{1}{\lambda_i}(h - \frac{n-p+1}{2\lambda_i})$, $h = \sum_{i=1}^p \frac{1}{\lambda_i}$.

Proof. We first provide the second fundamental form in explicit form. Using 4, we can compute and get

$$\boldsymbol{I\!I}_X(\boldsymbol{v}, \boldsymbol{v}) = Q\begin{pmatrix} 0_{p\times p} & 0_{p\times(n-p)} \\ 0_{(n-p)\times p} & \sum_{i=1}^p \frac{1}{\lambda_i}\boldsymbol{c}_i\boldsymbol{c}_i^T \end{pmatrix}Q^T \tag{3.4.26}$$

where $\boldsymbol{v} = \sum_{i=1}^p c_i\boldsymbol{v}_i + \sum_{i=1}^p\sum_{j=i+1}^n c_{ij}\boldsymbol{v}_{ij}$ for arbitrary coefficient $\{c_i\}_{i=1}^n \cup \{c_{ij}\}_{1\le i\le p, i<j\le n}$, and $\boldsymbol{c}_i = \sum_{j=p+1}^n c_{ij}e_j \in \mathbb{R}^{n-p}$ is a vector defined for $1 \le i \le p$.

Through parallelogram formula, we get

$$\boldsymbol{I\!I}_X(\boldsymbol{v}, \boldsymbol{v}') = Q\begin{pmatrix} 0_{p\times p} & 0_{p\times(n-p)} \\ 0_{(n-p)\times p} & \sum_{i=1}^p \frac{1}{\lambda_i}\frac{\boldsymbol{c}_i\boldsymbol{c}_i'^T + \boldsymbol{c}_i'\boldsymbol{c}_i^T}{2} \end{pmatrix}Q^T \tag{3.4.27}$$

By Gauss-Codazzi equation 16, we get the Riemann curvature tensor

$$\begin{aligned}R(\boldsymbol{v}',\boldsymbol{v},\boldsymbol{v},\boldsymbol{v}') =&\langle \boldsymbol{I\!I}_X(\boldsymbol{v},\boldsymbol{v}),\boldsymbol{I\!I}_X(\boldsymbol{v}',\boldsymbol{v}')\rangle - \langle \boldsymbol{I\!I}_X(\boldsymbol{v},\boldsymbol{v}'),\boldsymbol{I\!I}_X(\boldsymbol{v},\boldsymbol{v}')\rangle\\ =&\mathrm{Tr}\left((\sum_{j=1}^{p}\frac{1}{\lambda_{i'}}\boldsymbol{c}'_{i'}\boldsymbol{c}'^T_{i'})\times(\sum_{i=1}^{p}\frac{1}{\lambda_i}\boldsymbol{c}_i\boldsymbol{c}_i^T)\right) - \mathrm{Tr}\left((\sum_{i=1}^{p}\frac{1}{\lambda_i}\frac{\boldsymbol{c}_i\boldsymbol{c}'^T_i+\boldsymbol{c}'_i\boldsymbol{c}^T_i}{2})^2\right)\end{aligned} \tag{3.4.28}$$

Taking trace with respect to $\boldsymbol{v}'$, we get Ricci tensor as

$$\begin{aligned}\mathrm{Ric}(\boldsymbol{v},\boldsymbol{v}) =&\mathrm{Trace}_{v'}R(\boldsymbol{v}',\boldsymbol{v},\boldsymbol{v},\boldsymbol{v}') = \sum_{k=1}^{p}\sum_{j=p+1}^{n}R(\boldsymbol{v}_{kj},\boldsymbol{v},\boldsymbol{v},\boldsymbol{v}_{kj})\\ =&h\mathrm{Tr}\,(\sum_{i=1}^{p}\frac{1}{\lambda_i}\boldsymbol{c}_i\boldsymbol{c}_i^T) - \sum_{k=1}^{p}\sum_{j=p+1}^{n}\mathrm{Tr}\,(\frac{1}{2\lambda_k^2}\boldsymbol{c}_k\boldsymbol{c}'^T_k(\boldsymbol{c}_k\boldsymbol{c}'^T_k+\boldsymbol{c}'_k\boldsymbol{c}^T_k))\Big|_{\boldsymbol{c}'_k=e_j}\\ =&\sum_{i=1}^{p}h\times\frac{1}{\lambda_i}(\sum_{j=p+1}^{n}c_{ij}^2) - \sum_{k=1}^{p}\sum_{j=p+1}^{n}\frac{1}{2\lambda_k^2}(c_{kj}^2+|\boldsymbol{c}_k|^2)\\ =&\sum_{i=1}^{p}\frac{1}{\lambda_i}(h-\frac{n-p+1}{\lambda_i})(\sum_{j=p+1}^{n}c_{ij}^2)\end{aligned} \tag{3.4.29}$$

□

Remark 25. In fact, by taking the trace of Ric we can see that Ric is never positive definite when $p < \frac{n+1}{3}$, as

$$\mathrm{Tr}\,\mathrm{Ric} = \sum_{i=1}^{p}\sum_{j=p+1}^{n}\frac{1}{\lambda_i}(h-\frac{n-p+1}{2\lambda_i}) = (n-p)\Big((\sum_{i=1}^{p}\frac{1}{\lambda_i})^2 - \frac{n-p+1}{2}\sum_{i=1}^{p}\frac{1}{\lambda_i^2}\Big) \tag{3.4.30}$$

$$\leq (n-p)\frac{3p-n-1}{2}\sum_{i=1}^{p}\frac{1}{\lambda_i^2} \tag{3.4.31}$$

Corollary 5. *The Bakry-Émery-Ricci tensor* Ric_∞ *for* $(\mathcal{S}_+^{n,p}, g_E, \frac{1}{2}||X||_F^2)$ *is as fol-*

lows:

$$Ric_\infty(\boldsymbol{v},\boldsymbol{v}) = \sum_{i=1}^{p}(1+\frac{1}{\beta\lambda_i}(h-\frac{n-p+1}{\lambda_i}))(\sum_{j=p+1}^{n} c_{ij}^2) + \sum_{1\le i<j\le p} c_{ij}^2 \qquad (3.4.32)$$

with the smallest eigen value achieved at every vector in $\mathrm{Span}\{\boldsymbol{v}_{kj}\}_{j=p+1}^{n}$

$$r_{\infty,\min} = 1+\frac{1}{\beta\lambda_k}(h-\frac{n-p+1}{2\lambda_k}) > 1-\frac{n-p-1}{2\beta\lambda_p^2} \qquad (3.4.33)$$

where $k = \mathrm{argmin}_{i\in\{1,2,\dots,p\}}\frac{1}{\lambda_i}(h-\frac{n-p+1}{2\lambda_i})$*, and* λ_p *is the smallest eigenvalue of* X*.*

3.4.4 Bakry-Émery-Ricci tensor for the Bures-Wasserstein metric

In this section we consider the Bakry-Émery-Ricci tensor of the Bakry-Émery manifold $(\mathbb{R}_*^{n\times p}/\mathcal{O}_p, g_{BW}, h)$ where h is the energy function on $\mathbb{R}_*^{n\times p}/\mathcal{O}_p : h(\pi(Y)) = \mathcal{E}(YY^T)$. By O'Neill's formula [30] and flatness of $\mathbb{R}_*^{n\times p}$, the manifold $(\mathcal{S}_+^{n,p}, g_{BW})$ has nonnegative sectional curvature, therefore nonnegative Ricci tensor, which results in the following proposition.

Proposition 4. *For the Bakry-Émery manifold* $(\mathbb{R}_*^{n\times p}/\mathcal{O}_p, g_{BW}, h)$*, the Bakry-Émery-Ricci tensor satisfies the following inequality: for any* $\xi_{\pi(Y)} \in T_{\pi(Y)}\mathbb{R}_*^{n\times p}/\mathcal{O}_p$*,*

$$Ric_\infty(\xi_{\pi(Y)},\xi_{\pi(Y)}) \ge \left\langle \nabla^2\mathcal{E}(YY^T)[Y\overline{\xi}_Y^T+\overline{\xi}_YY^T], Y\overline{\xi}_Y^T+\overline{\xi}_YY^T\right\rangle + g_{BW}(2\nabla\mathcal{E}(YY^T)\overline{\xi}_Y, \overline{\xi_Y}). \qquad (3.4.34)$$

Proof. Recall that the Bakry-Émery Ricci tensor is defined as

$$\mathrm{Ric}_\infty = \frac{1}{\beta}\mathrm{Ric} + \mathrm{Hess}\,,$$

and recall that Ric is nonnegative as mentioned above, we have

$$\mathrm{Ric}_\infty \geq \mathrm{Hess}\,.$$

By [96], we have

$$g_{BW}(\mathrm{Hess}\, h(\pi(Y))[\xi_{\pi(Y)}], \xi_{\pi(Y)}) = \left\langle \nabla^2 \mathcal{E}(YY^T)[Y\overline{\xi}_Y^T + \overline{\xi}_Y Y^T], Y\overline{\xi}_Y^T + \overline{\xi}_Y Y^T \right\rangle + g_{BW}(2\nabla\mathcal{E}(YY^T)\overline{\xi}_Y, \overline{\xi_Y}). \tag{3.4.35}$$

Hence we completes the proof. $\square$

Corollary 6. *If in particular, $\mathcal{E}(X) = \frac{1}{2}\|X\|_F^2$, then* (3.4.34) *satisfies the following Bakry-Émery's curvature dimension condition.*

$$Ric_\infty(\xi_{\pi(Y)}, \xi_{\pi(Y)}) \geq 2\sigma_p g_{BW}(\xi_{\pi(Y)}, \xi_{\pi(Y)}), \quad \forall \xi_{\pi(Y)} \in T_{\pi(Y)}\mathbb{R}_*^{n\times p}/\mathcal{O}_p, \tag{3.4.36}$$

where $\sigma_p > 0$ is the smallest positive singular value of Y.

Proof. Since $\mathcal{E}(X) = \frac{1}{2}\|X\|_F^2$, we have

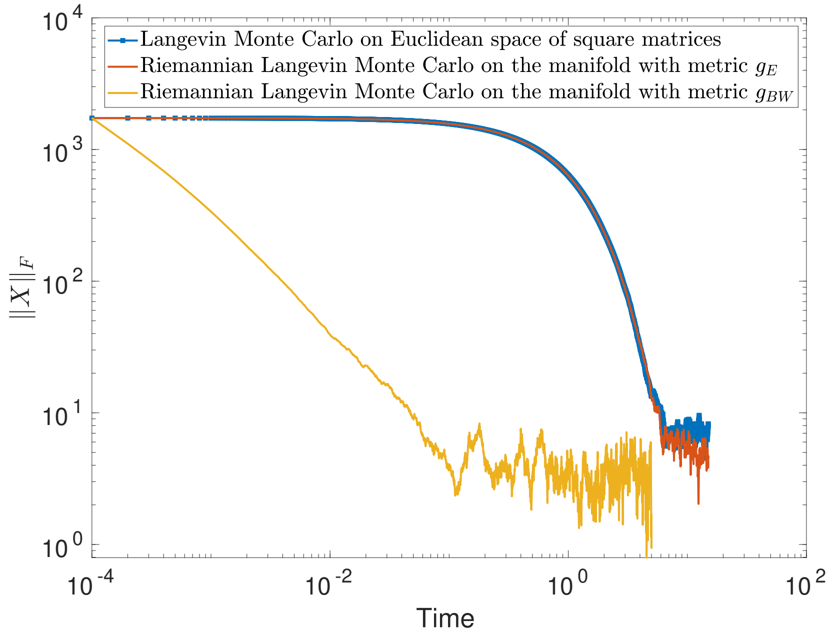
$$\begin{aligned}
\mathrm{Ric}_\infty(\xi_{\pi(Y)}, \xi_{\pi(Y)}) &\geq \left\langle \nabla^2 \mathcal{E}(YY^T)[Y\overline{\xi}_Y^T + \overline{\xi}_Y Y^T], Y\overline{\xi}_Y^T + \overline{\xi}_Y Y^T \right\rangle + g_{BW}(2\nabla\mathcal{E}(YY^T)\overline{\xi}_Y, \overline{\xi_Y}) \\
&= \left\| Y\overline{\xi}_Y^T + \overline{\xi}_Y Y^T \right\|_F^2 + \left\| Y^T \overline{\xi}_Y \right\|_F^2 \\
&\geq \left\| Y\overline{\xi}_Y^T + \overline{\xi}_Y Y^T \right\|_F^2 .
\end{aligned}$$

Then it follows from [96, Lemma 7.1] that $\left\| Y\overline{\xi}_Y^T + \overline{\xi}_Y Y^T \right\|_F^2 \geq 2\sigma_p g_{BW}(\xi_{\pi(Y)}, \xi_{\pi(Y)})$.

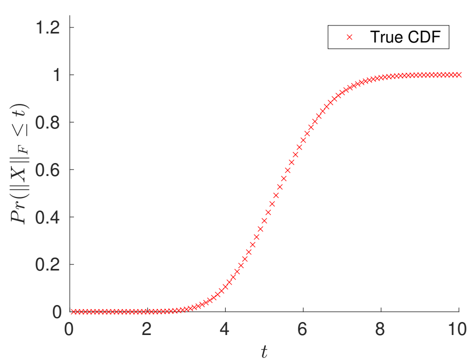
$\square$

3.4.5 Numerical illustration

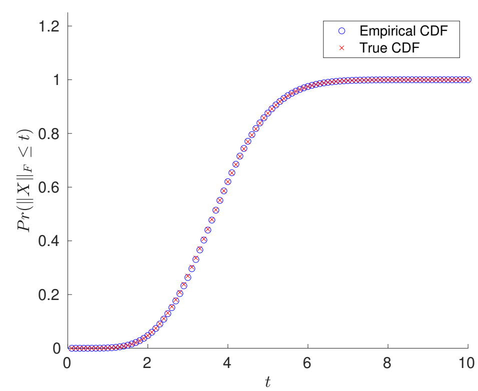
The mathematical theory of convergence speed of SDE to its equilibrium has been well studied, as is reviewed in [*cite analysis paper*]. One particular application of the two Riemannian Langevin schemes is to use them to numerically study the SDE solutions, e.g., by taking very small time steps, a Riemannian Langevin scheme approximates the Riemannian Langevin equation on the manifold. We have shown comparison of the Langevin equation on $(\mathcal{S}_+^{n,p}, g_E)$, $(\mathcal{S}_+^{n,p}, g_{BW})$, $\mathbb{R}^{n\times n}$ in Figure 3.4.1, in which we can see interesting differences between two metrics. With all three figures in Figure 3.4.1, we can see that the SDE on $(\mathcal{S}_+^{n,p}, g_{BW})$ has a much faster convergence to its Gibbs measure than the SDE on $(\mathcal{S}_+^{n,p}, g_E)$.



(a) Convergence of SDEs to the equilibrium.



(b) The true CDF of $D = \|X\|_F$ on $(\mathcal{S}_+^{n,p}, g_E)$. This implies the blue curve in Figure (a) has not reached equilibrium at Time=1.



(c) The true CDF of $D = \|X\|_F$ on $(\mathcal{S}_+^{n,p}, g_{BW})$, and empirical CDF of 5E6 samples from the Scheme BW after Time=5.

Figure 3.4.1: For $\mathcal{S}_+^{n,p}$ with $n = 5, p = 3$, the dimension is $N = 12$, and $\mathbb{R}^{n\times n}$ has dimension 25. The Gibbs measure is $e^{-\beta\mathcal{E}(x)}$ with $\beta = 0.4$ and $\mathcal{E} = \frac{1}{2}\|X\|_F^2$. The time step size is $1e-4$. The initial guess is a random PSD matrix of rank-3 with eigenvalues = [1000,1000,1000].

Chapter 4

Numerical schemes for Gibbs sampling problem

4.0 Abstract

This paper introduces two explicit schemes to sample matrices from Gibbs distributions on $\mathcal{S}_+^{n,p}$, the manifold of real positive semi-definite (PSD) matrices of size $n \times n$ and rank p. Given an energy function $\mathcal{E} : \mathcal{S}_+^{n,p} \to \mathbb{R}$ and certain Riemannian metrics g on $\mathcal{S}_+^{n,p}$, these schemes rely on an Euler-Maruyama discretization of the Riemannian Langevin equation (RLE) with Brownian motion on the manifold. We present numerical schemes for RLE under two fundamental metrics on $\mathcal{S}_+^{n,p}$: (a) the metric obtained from the embedding of $\mathcal{S}_+^{n,p} \subset \mathbb{R}^{n \times n}$; and (b) the Bures-Wasserstein metric corresponding to quotient geometry. We also provide examples of energy functions with explicit Gibbs distributions that allow numerical validation of these schemes.

4.1 Introduction

4.1.1 Problem statement

Consider the space of real, symmetric positive semi-definite matrices with size $n \times n$ and rank p, denoted by

$$\mathcal{S}_+^{n,p} = \{X \in \mathbb{R}^{n\times n} | X = X^T, X \succeq 0, \mathrm{rank}(X) = p\}. \tag{4.1.1}$$

Given an energy $\mathcal{E} : \mathcal{S}_+^{n,p} \to \mathbb{R}$ and a parameter $\beta > 0$ referred to as the inverse temperature, our goal is to sample efficiently from the Gibbs distribution

$$\rho_\beta(X) = \frac{1}{Z_\beta} e^{-\beta\mathcal{E}(X)} \rho_{\mathrm{ref}}(X), \quad Z_\beta = \int_{\mathcal{S}_+^{n,p}} e^{-\beta\mathcal{E}(X')} \rho_{\mathrm{ref}}(X')\, dX'. \tag{4.1.2}$$

Gibbs measures must be defined with respect to a base measure. In this work, we equip the space $\mathcal{S}_+^{n,p}$ with a Riemannian metric g and choose $\rho_{\mathrm{ref}}(X)dX = \sqrt{\det g(X)}dX$ to be the canonical volume form associated to the metric g. This volume form is expressed in coordinates for the metrics studied in this paper in Section 4.4.

This sampling problem is related to the optimization problem $\min_{X\in\mathcal{S}_+^{n,p}} \mathcal{E}(X)$ since in the limit $\beta \to \infty$ the Gibbs distribution concentrates at the global minima of $\mathcal{E}(X)$. Minimization problems over the space $\mathcal{S}_+^{n,p}$ arise in many areas, especially semidefinite programming and machine learning, and have been studied extensively. Gibbs distributions originate in statistical physics, while the sampling problem may also be seen as a stochastic variant of the optimization problem. For these reasons, the sampling problem has a broad range of applications; see Section 4.1.5 below.

The main contribution of this paper are efficient sampling schemes for ρ_β based on Langevin dynamics. Our approach builds on the geometric theory of optimization; in particular, we extend Riemannian optimization on $\mathcal{S}_+^{n,p}$ [91, 96] to Gibbs sampling as follows. In [91] it was recognized that two commonly used gradient descent schemes over $\mathcal{S}_+^{n,p}$ are time discretizations of *Riemannian* gradient flows, where $\mathcal{S}_+^{n,p}$ is equipped with the two natural Riemannian metrics listed below. We combine this observation with the theory of Brownian motion on Riemannian manifolds to obtain Riemannian Langevin equations and explicit sampling schemes.

The reader unfamiliar with these concepts should note that while the abstract theory serves to guide our work, the schemes presented in this paper may be implemented without requiring a complete understanding of the underlying theory. Further, while this paper is focused on the two numerical schemes below, the underlying framework can be used to extend other Riemannian gradient descent schemes to sampling schemes for the Gibbs measure. The new phenomenon that arises is the interplay between Brownian motion and curvature in the Riemannian Langevin equation. This interplay has been studied in depth by two of the authors (TY and GM) and their co-workers in recent papers for geometries used in optimization and physics [48, 71].

4.1.2 Two Riemannian metrics on $\mathcal{S}_+^{n,p}$

Given $X \in \mathcal{S}_+^{n,p}$, let $X = YY^T$ be a low-rank decomposition where $Y \in \mathbb{R}^{n\times p}$. We use two fundamental metrics on $\mathcal{S}_+^{n,p}$ obtained from this parametrization, from the Euclidean metric for either the variable X or the variable Y through the use of Riemannian embedding and Riemannian submersion respectively. These are the two most natural ways of defining metrics on $\mathcal{S}_+^{n,p}$.

The flat metric for X corresponds to the embedded geometry of $\mathcal{S}_+^{n,p}$ in the Euclidean space $\mathbb{R}^{n\times n}$ [91]. Precisely, we consider the natural Riemannian embedding $\mathcal{S}_+^{n,p} \hookrightarrow \mathbb{R}^{n\times n}$ and use the Frobenius norm on $\mathbb{R}^{n\times n}$ to define a metric on $\mathcal{S}_+^{n,p}$. Denote it by g_E, then $g_E(A,B) = \mathrm{Tr}(A^TB)$ for any two square matrices A, B in the tangent space of $\mathcal{S}_+^{n,p}$, where Tr denotes the trace of a matrix.

On the other hand, we may also use the flat geometry on Y to define a metric on $\mathcal{S}_+^{n,p}$. We observe that if $YY^T = X$, then it is also true that $\tilde{Y}\tilde{Y}^T = X$ where $\tilde{Y} = YO$ and $O \in \mathcal{O}_p$, the orthogonal group of dimension p. Thus, we may identify $\mathcal{S}_+^{n,p} \approx \mathbb{R}_*^{n\times p}/\mathcal{O}_p$, as a quotient space, with a quotient map

$$\begin{aligned} \pi : \mathbb{R}_*^{n\times p} &\to \mathbb{R}_*^{n\times p}/\mathcal{O}_p \\ Y &\mapsto [Y] = \{YO \mid O \in \mathcal{O}_p\}. \end{aligned}$$

Here $\mathbb{R}_*^{n\times p}$ denotes full rank matrices.

The quotient space structure can be enhanced with a Riemannian metric through the use of Riemannian submersion. Roughly, the metric for X corresponds to the metric for Y in a manner that respects the splitting of the tangent space at Y into the space of the group action and its complement. If $\mathbb{R}_*^{n\times p}$ is equipped with Euclidean metric, then the metric induced by the submersion is often called the Bures-Wasserstein metric on $\mathcal{S}_+^{n,p} \approx \mathbb{R}_*^{n\times p}/\mathcal{O}_p$, denoted by g_{BW} (see [13, 69, 70]).

4.1.3 Langevin dynamics and the Riemannian Langevin equation

We now explain how Langevin equations may be defined intrinsically on $(\mathcal{S}_+^{n,p}, g)$.

Let us first recall the Langevin equation on $\mathbb{R}^n$. Assume given a potential or energy function $\mathcal{E} : \mathbb{R}^n \to \mathbb{R}$ and let W_t denote the standard Wiener process on $\mathbb{R}^n$. The Langevin equation for the potential $\mathcal{E}$ is the Itô differential equation

$$dx_t = -\nabla\mathcal{E}(x_t)\, dt + \sqrt{\frac{2}{\beta}}\, dW_t. \tag{4.1.3}$$

The Fokker-Planck equation describes the evolution of the probability density of x_t.

With $\rho(x,t)\,dx = \mathbb{P}(x_t \in (x, x+dx))$, we have

$$\partial_t \rho = \frac{1}{\beta}\triangle\rho + \nabla\cdot(\rho\nabla\mathcal{E})\,. \tag{4.1.4}$$

The Gibbs density (with reference density being uniform with respect to Lebesgue measure) is the unique equilibrium of equation (4.1.4) under natural growth assumptions on the energy $\mathcal{E}$ as $|x| \to \infty$.

The Langevin equation immediately yields a numerical scheme for (approximate) sampling from the Gibbs distribution. Fix a step size $\Delta t > 0$, let $t_k = k\Delta t$, $k = 0, 1, \ldots$, and let x_k denote the numerical approximation to (4.1.3) at time t_k. The Euler-Maruyama scheme to approximate equation (4.1.3), also known as Langevin Monte Carlo in the statistics literature, is

$$x_{k+1} = x_k - \Delta t\nabla\mathcal{E}(x_k) + \sqrt{\frac{2\Delta t}{\beta}}\xi_k, \tag{4.1.5}$$

where $\xi_k = (\xi_k^1, \ldots, \xi_k^n)$ is an i.i.d. sequence of standard Gaussian vectors in $\mathbb{R}^n$. This scheme is explicit. In order to extend it to sampling from (4.1.2) we must understand how to modify the Langevin equation on the Riemannian manifold $(\mathcal{S}_+^{n,p}, g)$.

First, the term $\nabla\mathcal{E}$ must be replaced by the Riemannian gradient, written as $\mathrm{grad}\mathcal{E}$. The more subtle modification of equation (4.1.3) concerns the noise. The natural analogy is to replace the Wiener process W_t on $\mathbb{R}^n$ with Brownian motion on the Riemannian manifold $(\mathcal{S}_+^{n,p}, g)$ at inverse temperature β, denoted $\boldsymbol{B}_t^{g,\beta}$. This yields the (formal) Riemannian Langevin equation on $(\mathcal{S}_+^{n,p}, g)$

$$\mathrm{d}\boldsymbol{X}_t = -\mathrm{grad}\mathcal{E}(\boldsymbol{X}_t)\mathrm{d}t + \mathrm{d}\boldsymbol{B}_t^{g,\beta}. \tag{4.1.6}$$

This equation is only formal because stochastic differential equations on manifolds must be defined using the Stratonovich formulation in order to ensure coordinate independence (Itô differentials do not satisfy the chain rule, while Stratonovich differentials do) [43, 47]. On the other hand, Itô differential equations are convenient for analysis as well as simulation. Thus, in formulating the Riemannian Langevin equation, it is necessary to first formulate the appropriate Stratonovich equation and then compute the deterministic Itô–Stratonovich correction. A central observation in our work is that this correction term is due to curvature and is explicitly computable for several Riemannian geometries relevant to optimization [44, 48, 71].

4.1.4 Riemannian Langevin Monte Carlo sampling schemes

For the two metrics considered in this paper, the Itô–Stratonovich correction due to curvature may also be computed explicitly, yielding the SDEs in Section 4.2. The rigorous analysis of these SDEs is presented in the companion paper [93], and we focus on numerical algorithms in this paper. The Euler-Maruyama approximation to these SDEs yields the numerical sampling schemes listed below.

The SDEs also admit other numerical approximations. We have chosen the Euler-Maruyama schemes because these schemes are fully explicit, simple to state, implement and numerically validate. They are generalizations of the popular unadjusted Langevin Monte Carlo for sampling in Euclidean spaces. Further, these schemes reduce to deterministic Riemannian gradient descent methods in the limit $\beta \to \infty$.

Scheme E for the embedded geometry

For the embedded manifold $(\mathcal{S}_+^{n,p}, g_E)$, the scheme is

$$X_{k+1} = \mathrm{P}_{\mathcal{S}_+^{n,p}}\left[X_k - \Delta t\,\mathrm{grad}\mathcal{E}(X_k) + Q_k\left(\sqrt{\frac{2\Delta t}{\beta}}\begin{bmatrix} B_{11} & B_{12} \\ B_{12}^T & 0 \end{bmatrix} + \frac{\Delta t}{\beta}\sum_{i=1}^{p}\tfrac{1}{\lambda_i}\begin{bmatrix} 0 & 0 \\ 0 & I_{n-p} \end{bmatrix}\right)Q_k^T\right], \tag{4.1.7}$$

where $\mathrm{P}_{\mathcal{S}_+^{n,p}}$ is the Euclidean projection to $\mathcal{S}_+^{n,p}$, and $X_k = Q_k \Lambda Q_k^T$ is the full SVD of $X_k \in \mathcal{S}_+^{n,p}$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p > 0$. The entries of B_{12} are i.i.d. drawn from $\sqrt{\frac{1}{2}}\mathcal{N}(0,1)$. The entries of the symmetric B_{11} are defined as follows: the diagonal entries are i.i.d. drawn from $\mathcal{N}(0,1)$, and off-diagonal entries are $b_{ij} = b_{ji} \sim \sqrt{\frac{1}{2}}\mathcal{N}(0,1)$. When $\beta = \infty$, equation (4.1.7) reduces to $X_{k+1} = \mathrm{P}_{\mathcal{S}_+^{n,p}}(X_k - \Delta t \operatorname{grad}\mathcal{E}(X_k))$, which is the Riemannian gradient descent on $(\mathcal{S}_+^{n,p}, g_E)$, see [1, 96]. We refer to (4.1.7) as Scheme E.

In this scheme, the term $\dfrac{\Delta t}{\beta}\sum\limits_{i=1}^{p}\frac{1}{\lambda_i}\begin{bmatrix} 0 & 0 \\ 0 & I_{n-p}\end{bmatrix}$ in equation (4.1.7) is the correction due to the mean curvature of the embedding of $\mathcal{S}_+^{n,p} \hookrightarrow \mathbb{R}^{n\times n}$.

Scheme BW for the Bures-Wasserstein metric

For the quotient manifold $(\mathbb{R}_*^{n\times p}/\mathcal{O}_p, g_{BW})$, the scheme is

$$Y_{k+1} = Y_k - \Delta t 2\nabla\mathcal{E}(Y_k Y_k^T)Y_k + \sqrt{\frac{2\Delta t}{\beta}}B_k + \frac{\Delta t}{\beta}U_k \left[\sum_{j:j\neq i}\tfrac{\sigma_i}{\sigma_i^2+\sigma_j^2}\right]_{ii} V_k^T, \tag{4.1.8}$$

where B_k is n-by-p matrix with entries being i.i.d. standard Gaussian, $Y_k = U_k\Sigma_k V_k^T \in \mathbb{R}^{n\times p}$ is the compact SVD with singular values σ_i, and $\left[\sum\limits_{j:j\neq i}\tfrac{\sigma_i}{\sigma_i^2+\sigma_j^2}\right]_{ii}$ is the diagonal matrix whose i-th diagonal entry is $\sum\limits_{j:j\neq i}\tfrac{\sigma_i}{\sigma_i^2+\sigma_j^2}$. We refer to (4.1.8) as Scheme BW. The Riemannian Langevin Monte Carlo scheme (4.1.8) can be viewed as a natural extension of Burer-Monteiro gradient descent method

$$Y_{k+1} = Y_k - \Delta t\, 2\nabla\mathcal{E}(Y_k Y_k^T)Y_k, \tag{4.1.9}$$

which is the simplest low-rank gradient descent method for minimizing $\mathcal{E}(X)$ under the constraint $X \in \mathcal{S}_+^{n,p}$. It is clear that as $\beta \to \infty$, (4.1.8) reduces to (4.1.9). The Burer-Monteiro gradient descent method is equivalent to a Riemannian gradient descent method on the quotient manifold $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$ with Bures-Wasserstein metric, see [96].

Gibbs distribution sampling and numerical validation

While the Gibbs distribution always has the same density function $e^{-\beta\mathcal{E}}$ with respect to ρ_{ref}, the reference density ρ_{ref} depends on the metric. Thus, the two schemes (4.1.7) and (4.1.8), generate samples for two different probability distributions. In order to validate our schemes, we choose energy functions that allow an explicit computation of these densities for both metric. These energy functions yield matrix integrals of independent analytic interest. They also allow side-to-side benchmarking for different Gibbs samplers on $\mathcal{S}_+^{n,p}$. We demonstrate the efficiency of sampling from these Gibbs distributions numerically. Further analysis on convergence to equilibrium as $t \to \infty$ using the Bakry-Emery criterion is considered in the companion paper [93].

Finally, while we do not discuss their convergence and efficiency approximating SDE as the step-size $\Delta t \to 0$; this is possible following existing approximation results [56, 24, 62].

4.1.5 Some applications and related work

Applications of PSD matrices

Positive semi-definite (PSD) fixed rank matrices arise in many problems such as distance matrices [88] and covariance matrices in statistics, and have been used in applications including kernels in machine learning [73], semidefinite optimization [19], quantum information, etc. Riemannian optimization algorithms over $\mathcal{S}_+^{n,p}$ under different metrics have been well studied, e.g., see [91, 46, 70, 96] and references therein.

Langevin dynamics and Monte Carlo schemes on manifolds

There is an extensive literature on Langevin dynamics in statistics and related areas, with interest in nonconvex optimization [22, 23], as well as machine learning such as generative models [31].

In recent years, there has been interest in studying Langevin diffusion and Monte Carlo Markov Chain (MCMC) schemes on manifolds [25, 26, 38, 17, 20, 95, 74, 36, 58, 62]. In this paper, we are interested in Riemannian Langevin Monte Carlo schemes on $\mathcal{S}_+^{n,p}$.

In the statistics literature, manifold Langevin schemes have been studied in [38, 20]. However, these schemes apply only to simpler embedded manifolds $\mathcal{M} \subset \mathbb{R}^n$ with explicit geodesics such as the sphere and Stiefel manifolds. The above schemes do not directly apply to the manifold $\mathcal{S}_+^{n,p}$, even for the embedded geometry. In [95], a sampling scheme using projection to surface is constructed; however, this is not a Langevin scheme.

In general, a Langevin scheme can be used for either optimization [92, 62], or Monte Carlo type numerical integration, which is common in Bayesian statistic. For optimization, stochastic optimization by Langevin dynamics with simulated annealing is an established approach [65]. In [22], underdamped Langevin schemes are shown to be much more efficient than the overdamped case (4.1.5). For sampling, Metropolis-adjusted Langevin algorithm [38] is often used. For simplicity, we focus on the simple schemes (4.1.7) and (4.1.8) without considering any of simulated annealing, underdamped Langevin, or Metropolis-adjustment, to which it is possible to extend our schemes. Though the Riemannian optimization on $\mathcal{S}_+^{n,p}$ can be easily extended to Hermitian PSD matrices of fixed rank [96], we remark that such an extension for Langevin dynamics would be significantly different.

4.1.6 Organization of the paper

In Section 4.2, we state the explicit formulae for the SDE (4.1.6) and Gibbs measure on the manifold $\mathcal{S}_+^{n,p}$ under two metrics g_E and g_{BW}. We then derive the schemes (4.1.7) and (4.1.8) in Section 4.3. The energy functions and Gibbs distributions used to benchmark the schemes are presented in Section 4.4. The numerical results are studied in Section 4.5.

4.2 Riemannian Langevin equations on $\mathcal{S}_+^{n,p}$

In this section, we state the Itô form of the Riemannian Langevin equation (4.1.6) for both Riemannian geometries studied in this paper. The theoretical basis for these SDEs is discussed at greater depth in [93]. The main ideas are as follows: (a) the abstract theory of Brownian motion on Riemannian manifolds is used to define the Riemannian Langevin equation in Stratonovich form for the metrics g_E and g_{BW} on $\mathcal{S}_+^{n,p}$; (b) the Itô-Stratonovich conversion rule is used to compute the associated Itô form of these SDEs and it is observed that the Itô-Stratonovich correction term corresponds to mean curvature. This approach yields the SDEs below. These SDEs are used to develop numerical schemes in Section 4.3.

4.2.1 The Riemannian Langevin equation for embedded geometry $(\mathcal{S}_+^{n,p}, g_E)$

Let $X \in S_+^{n,p}$ have the compact SVD $X = U\Lambda U^T$ with $U \in \mathbb{R}^{n\times p}$. Let $U_\perp \in \mathbb{R}^{n\times(n-p)}$ be a matrix with columns orthonormal to columns of U. The tangent space of $S_+^{n,p}$ at $X = U\Lambda U^T \in S_+^{n,p}$ is given by [91, 96]:

$$T_X S_+^{n,p} = \left\{ \begin{bmatrix} U & U_\perp \end{bmatrix} \begin{bmatrix} H & K^T \\ K & 0 \end{bmatrix} \begin{bmatrix} U^T \\ U_\perp^T \end{bmatrix} : \forall K \in \mathbb{R}^{(n-p)\times p}, \forall H \in \mathbb{R}^{p\times p}, H^T = H \right\}. \tag{4.2.1}$$

The induced metric g_E by the embedding $\mathcal{S}_+^{n,p} \hookrightarrow \mathbb{R}^{n\times n}$ is then defined as

$$g_E(A, B) = \mathrm{Tr}(A^T B), \quad \forall A, B \in T_X S_+^{n,p},$$

which is the Frobenius inner product for two matrices.

Equation (4.1.6) describes the evolution of a point $\boldsymbol{X}_t \in \mathcal{S}_+^{n,p}$ in abstract terms.

We now rewrite it in a simpler equivalent form describing the evolution of the entries of the matrix entries $\{(X_t)_{ij}\}_{i,j=1}^n$ representing $\boldsymbol{X}_t$. Let us write $X = U\Lambda U^T$ for the compact singular value decomposition (SVD) of X. We further assume that the singular values $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_p)$ are written in decreasing order. We suppress the subscript t in the following equations, though the reader should note that U and Λ depend on X_t.

Then we find that the law of X_t is determined by the Itô differential equation

$$\mathrm{d}X_t = -\mathrm{grad}\mathcal{E}(X_t)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}W_t^{n,p,X_t} + \frac{1}{\beta}H(X_t)\mathrm{d}t. \tag{4.2.2}$$

In this equation, the stochastic forcing W_t^{n,p,X_t} is the orthogonal projection of white noise in $\mathbb{R}^{n\times n}$ onto $T_{X_t}\mathcal{S}_+^{n,p}$. Precisely, given W_t^i for $1 \le i \le n$ and $W_t^{i,j}$ for $1 \le i < j \le n$ independent standard one-dimensional Wiener process, we set

$$\mathrm{d}W_t^{n,p,X_t} = \begin{bmatrix} U & U_\perp \end{bmatrix} \begin{bmatrix} \mathrm{d}W_t^1 & \cdots & \frac{1}{\sqrt{2}}\mathrm{d}W_t^{1,p} & \frac{1}{\sqrt{2}}\mathrm{d}W_t^{1,p+1} & \cdots & \frac{1}{\sqrt{2}}\mathrm{d}W_t^{1,n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{2}}\mathrm{d}W_t^{1,p} & \cdots & \mathrm{d}W_t^p & \frac{1}{\sqrt{2}}\mathrm{d}W_t^{p,p+1} & \cdots & \frac{1}{\sqrt{2}}\mathrm{d}W_t^{p,n} \\ \frac{1}{\sqrt{2}}\mathrm{d}W_t^{1,p+1} & \cdots & \frac{1}{\sqrt{2}}\mathrm{d}W_t^{p,p+1} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{2}}\mathrm{d}W_t^{1,n} & \cdots & \frac{1}{\sqrt{2}}\mathrm{d}W_t^{p,n} & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} U^T \\ U_\perp^T \end{bmatrix},$$

The term $H(X_t)$ is the mean curvature of the embedding $\mathcal{S}_+^{n,p} \to \mathbb{R}^{n\times n}$. We adopt the convention in geometric analysis: the mean curvature is defined as the trace of the second fundamental form of the embedding. Explicitly, we have

$$H(X_t) = \left(\sum_{i=1}^p \frac{1}{\lambda_i}\right) \begin{bmatrix} U & U_\perp \end{bmatrix} \begin{bmatrix} 0_{p\times p} & 0_{p\times(n-p)} \\ 0_{(n-p)\times p} & I_{n-p} \end{bmatrix} \begin{bmatrix} U^T \\ U_\perp^T \end{bmatrix}. \tag{4.2.3}$$

The following feature of equation (4.2.2) is fundamental. The stochastic forcing is the naive projection of white noise in the ambient space $\mathbb{R}^{n\times n}$ onto $T_{X_t}\mathcal{S}_+^{n,p}$. Intuitively, when one uses the Euler-Maruyama discretization, the role of this term is to update X_t by taking unbiased random steps in any direction in the tangent space. However, Itô calculus has a subtle interplay with the geometry of the embedding, and in order to keep X_t on the manifold $\mathcal{S}_+^{n,p}$, it is necessary to include the correction term given by the mean curvature.

4.2.2 The Riemannian Langevin equation for Bures-Wasserstein geometry $(\mathcal{S}_+^{n,p}, g_{BW})$

The manifold $S_+^{n,p}$ can also be viewed as a quotient manifold $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$, for which the noncompact Stiefel manifold $\mathbb{R}_*^{n\times p}$ is called the *total space.* Denote the natural projection as

$$\pi : \mathbb{R}_*^{n\times p} \to \mathbb{R}_*^{n\times p}/\mathcal{O}_p.$$

For any $Y \in \mathbb{R}_*^{n\times p}$, the equivalence class containing Y is

$$[Y] = \pi^{-1}(\pi(Y)) = \{YO \mid O \in \mathcal{O}_p\},$$

which is an embedded submanifold of $\mathbb{R}_*^{n\times p}$ (see e.g., [1, Prop. 3.4.4]). The tangent space of $[Y]$ at Y is therefore a subspace of $T_Y\mathbb{R}_*^{n\times p}$ called the *vertical space* at Y, and is denoted by $\mathcal{V}_Y = \left\{Y\Omega \mid \Omega^T = -\Omega, \Omega \in \mathbb{R}^{p\times p}\right\}$, see [96].

Define

$$\begin{aligned} \theta : \mathbb{R}_*^{n\times p} &\to \mathcal{S}_+^{n,p} \\ Y &\mapsto YY^T. \end{aligned}$$

Then θ is invariant under the equivalence relation and induces a bijection $\tilde{\theta}$ on $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$ such that $\theta = \tilde{\theta} \circ \pi$. For any function $\mathcal{E}(X)$ defined on $\mathcal{S}_+^{n,p}$, there is a function F defined on $\mathbb{R}_*^{n\times p}$ that induces $\mathcal{E}$: for any $X = YY^T \in \mathcal{S}_+^{n,p}$, $F(Y) :=$

$\mathcal{E} \circ \theta(Y) = \mathcal{E}(YY^T)$. This is summarized in the diagram below:

$$\begin{tikzcd} \mathbb{R}_*^{n\times p} \arrow[d, "\pi"] \arrow[dr, dashed, "\theta:=\tilde{\theta}\circ\pi"] & & \\ \mathbb{R}_*^{n\times p}/\mathcal{O}_p \arrow[r, leftrightarrow, "\tilde{\theta}"] & \mathcal{S}_+^{n,p} \arrow[r, "\mathcal{E}"] & \mathbb{R} \end{tikzcd}$$

In particular, $\mathcal{S}_+^{n,p}$ is diffeomorphic to $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$ under $\tilde{\theta}$, see [96]. For any $Y \in \mathbb{R}_*^{n\times p}$, the flat metric for the total space $\mathbb{R}_*^{n\times p}$, correction term corresponds to mean curvature.

$$g(a, b) = \mathrm{Tr}(a^T b), \forall a, b \in T_Y \mathbb{R}_*^{n\times p} = \mathbb{R}^{n\times p}$$

induces a metric on the quotient manifold $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$, which is called Bures-Wasserstein metric, see [70, 69, 96]. Another way to understand the Bures-Wasserstein metric at $X \in \mathcal{S}_+^{n,p} \approx \mathbb{R}_*^{n\times p}/\mathcal{O}_p$ is via the map θ:

$$\begin{aligned} g_{BW}(A, B) = \mathrm{Tr}(ab^T) \quad & \forall A, B \in T_X\mathcal{S}_+^{n,p}, a, b \in T_Y\mathbb{R}_*^{n\times p} \\ & \text{s.t. } \mathrm{d}\theta(Y)[a] = A, \mathrm{d}\theta(Y)[b] = B, a, b \in \mathrm{Ker}(\mathrm{d}\theta(Y))^\perp \end{aligned} \tag{4.2.4}$$

where X has decomposition $X = YY^T$, $\mathrm{d}\theta(Y)[a] = Ya^T + aY^T$ is the differential of θ at Y, and $a \in \mathrm{Ker}(\mathrm{d}\theta(Y))^\perp \Leftrightarrow Y^T a = a^T Y$.

The Riemannian Langevin equation is now determined by the geometry of Riemannian submersion. We must obtain an Itô differential equation for Y_t, such that $X_t = Y_t Y_t^T$ is a matrix that has the same law as the solution to (4.1.6) in $(\mathcal{S}_+^{n,p}, g_{BW})$.

In comparison with equation (4.2.2), we see that the natural choice for white noise driving Y_t is white noise in $\mathbb{R}^{n\times p}$. This is the stochastic differential $\mathrm{d}W_t$, where $W_t = \{W_t^{ij}\}_{1\le i\le n, 1\le j\le p}$ consists of np independent standard one-dimensional Wiener processes. However, as in equation (4.2.2) we must include a deterministic correction. This correction corresponds to mean curvature again, but in a more subtle way than (4.2.2). The equivalence class of Y such that $X = YY^T$ is a group orbit of $\mathcal{O}_p$ embedded within $\mathbb{R}^{n\times p}$. The logarithm of the volume of this group orbit constitutes a natural Boltzmann entropy. It may be computed explicitly, and we find

$$S(Y) = \frac{1}{2}\sum_{i=1}^{p}\sum_{j=i+1}^{p}\log(\sigma_i^2+\sigma_j^2) \tag{4.2.5}$$

where $\{\sigma_i\}_{i=1}^p$ are singular values of Y. It is known that $\nabla S(Y)$ is the mean curvature of the group orbit in $\mathbb{R}^{n\times p}$ [78, p.3505].

We then have the following Itô differential equation for Y_t such that $X_t = Y_t Y_t^T$ has the same law as the solution to (4.1.6).

$$\mathrm{d}Y_{ij} = -\frac{\partial \mathcal{E}(YY^T)}{\partial Y_{ij}}\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}W_t^{ij} - \frac{1}{\beta}\frac{\partial S(Y)}{\partial Y_{ij}}\mathrm{d}t, \quad 1\le i\le n, 1\le j\le p. \tag{4.2.6}$$

The correction term can be explicitly computed using the following

Lemma 18. *If $Y \in R_*^{n\times p}$ has SVD as $Y = Q\Sigma P^T$ with singular values σ_i, then the gradient of the correction term S is given by $\nabla S(Y) = Q\tilde{\Sigma}P^T$ where $\tilde{\Sigma}$ is a diagonal matrix with diagonal entries $\sum_{j\neq 1}\frac{\sigma_1}{\sigma_1^2+\sigma_j^2}, \sum_{j\neq 2}\frac{\sigma_2}{\sigma_2^2+\sigma_j^2}, \cdots, \sum_{j\neq p}\frac{\sigma_p}{\sigma_p^2+\sigma_j^2}$.*

4.3 Two Riemannian Langevin Monte Carlo schemes

To get a simple Riemannian Langevin Monte Carlo sampling scheme, we only consider convenient discretization and approximation methods, which can be easily and efficiently implemented. For the Brownian motion term, we consider the most straightforward and simplest discretization of the SDEs (4.2.2) and (4.2.6), i.e., the Euler-Maruyama type discretization.

One extra complication from the manifold constraint is how to approximate the exponential map. For optimization algorithms on Riemannian manifolds [1], retraction, which is at least a first order approximation to the exponential map, is often used. For instance, for approximating an ODE $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{X} = -\mathrm{grad}\mathcal{E}(\boldsymbol{X})$ on a manifold $\mathcal{M}$, with any retraction operator $\mathcal{R}_{\mathcal{M}}$ mapping to $\mathcal{M}$, a simple forward Euler type approximation, or equivalently the Riemannian gradient descent method, is given by

$$\boldsymbol{X}_{k+1} = \mathcal{R}_{\mathcal{M}}[\boldsymbol{X}_{k+1} - \Delta t \, \mathrm{grad}\mathcal{E}(\boldsymbol{X}_k)].$$

In particular, when combining the Euler-Maruyama type discretization for SDE and the simple Riemannian gradient descent by retraction, we get the two simple Riemmanian Langevin Monte Carlo schemes as follows.

4.3.1 Scheme E for the embedded geometry

The Riemannian gradient

For a given energy function $\mathcal{E}(X)$, its Riemannian gradient $\mathrm{grad}\mathcal{E}(X)$ of at $X \in \mathcal{S}_+^{n,p}$, is the Euclidean projection of the Euclidean gradient $\nabla\mathcal{E}(X) \in \mathbb{R}^{n\times n}$ defined as $[\nabla\mathcal{E}(X)]_{ij} = \frac{\partial}{\partial X_i}\mathcal{E}(X)$, onto the tangent space $T_X\mathcal{S}_+^{n,p}$, see [1, 91, 96]. It is straightforward to verify that $\nabla\mathcal{E}(X)$ is a symmetric matrix for any differentiable $\mathcal{E}$ and any $X \in \mathcal{S}_+^{n,p}$. For any given $X \in \mathcal{S}_+^{n,p}$, let $X = U\Lambda U^T$ be its compact SVD. Let $P_U = UU^T$ and $P_{U_\perp} = U_\perp U_\perp^T = I - UU^T$. By derivations in [96], $\mathrm{grad}\mathcal{E}(X)$ can be computed and represented as

$$\begin{aligned}\mathrm{grad}\mathcal{E}(X) &= \begin{bmatrix} U & U_\perp \end{bmatrix} \begin{bmatrix} U^T\nabla\mathcal{E}(X)U & U^T\nabla\mathcal{E}(X)U_\perp \\ U_\perp^T\nabla\mathcal{E}(X)U & 0 \end{bmatrix} \begin{bmatrix} U^T \\ U_\perp^T \end{bmatrix} \\ &= P_U\nabla\mathcal{E}(X)P_U + P_{U_\perp}\nabla\mathcal{E}(X)P_U + P_U\nabla\mathcal{E}(X)P_{U_\perp}.\end{aligned}$$

The compact implementation of computing $\mathrm{grad}\mathcal{E}(X)$ is given in Algorithm 1.

Algorithm 1 Compact computation of the Riemannian gradient $\mathrm{grad}\mathcal{E}(X)$

Require: The compact SVD of $X \in S_+^{n,p}$: $X = U\Lambda U^T$

Ensure: $\mathrm{grad}\mathcal{E}(X) = UHU^T + U_pU^T + UU_p^T \in T_X S_+^{n,p}$

$T \leftarrow \nabla\mathcal{E}(X)U$

$H \leftarrow U^T T$

$U_p \leftarrow T - UH$

The retraction by projection

Let $\mathcal{S}^{n\times n}$ denote symmetric matrices, then the Euclidean projection $\mathrm{P}_{S_+^{n,p}} : \mathcal{S}^{n\times n} \longrightarrow \mathcal{S}_+^{n,p}$ is a convenient retraction operator, see [1, 91, 96]. A straightforward implementation is given in Aglorithm 2.

Algorithm 2 Computation of the retraction $P_{S_+^{n,p}}(X+Z)$

Require: the compact SVD of X: $X = U\Lambda U^T \in S_+^{n,p}$, $Z \in \mathcal{S}^{n\times n}$.

Ensure: $P_{S_+^{n,p}}(X+Z) = Q_+\Lambda_+Q_+^T \in S_+^{n,p}$.

$(Q_+, \Lambda_+) = \mathrm{svd}(X+Z)$

$U_+ \leftarrow Q_+(:, 1:p) \quad \Lambda_+ \leftarrow \Lambda_+(1:p, 1:p)$

A Riemannian Langevin Monte Carlo scheme

For approximating the SDE (4.2.2) on $(S_+^{n,p}, g_E)$, with the retraction operator and Euler-Maruyama method for SDE, we have the scheme (4.1.7), which can be more explicitly written as

$$X_{k+1} = \mathrm{P}_{S_+^{n,p}}\left(\begin{bmatrix} U & U_\perp \end{bmatrix} \begin{bmatrix} \Lambda - \Delta t U^T \nabla \mathcal{E}(X_k) U + \sqrt{\frac{2\Delta t}{\beta}} B_{11} & -\Delta t U^T \nabla \mathcal{E}(X_k) U_\perp + \sqrt{\frac{2\Delta t}{\beta}} B_{12} \\ -\Delta t U_\perp^T \nabla \mathcal{E}(X_k) U + \sqrt{\frac{2\Delta t}{\beta}} B_{12}^T & \frac{\Delta t}{\beta} \sum_{i=1}^{p} \frac{1}{\lambda_i} I_{n-p} \end{bmatrix} \begin{bmatrix} U^T \\ U_\perp^T \end{bmatrix}\right), \tag{4.3.1}$$

where $X_k = U \Lambda U^T$ is the compact SVD of $X_k \in S_+^{n,p}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. The third term in the right hand side is the white noise term in the tangent space $T_{X_k}\mathcal{S}_+^{n,p}$. Entries of $B_{12} \in \mathbb{R}^{p\times(n-p)}$ are i.i.d drawn from $\sqrt{\frac{1}{2}}\mathcal{N}(0,1)$, and $B_{11} \in \mathbb{R}^{p\times p}$ are defined as follows.

$$B_{11} = \begin{bmatrix} \mathcal{N}(0,1) & & & \\ & \ddots & b_{ij} & \\ & b_{ji} & \ddots & \\ & & & \mathcal{N}(0,1) \end{bmatrix} \tag{4.3.2}$$

with $b_{ij} = b_{ji} \sim \sqrt{\frac{1}{2}}\mathcal{N}(0,1)$. The implementation details of the scheme (4.1.7) are given as follows in the Algorithm 3.

Algorithm 3 The Riemannian Langevin Monte Carlo scheme (4.1.7) for $(\mathcal{S}_+^{n,p}, g_E)$

Require: initial iterate $X_1 \in \mathcal{S}_+^{n,p}$; full SVD of X_1: $X_1 = Q_1 \Lambda_1 Q_1^T$

1: **for** $k = 1, 2, \ldots, N$ **do**

2: Compute Riemannian gradient

$$\xi_k := \mathrm{grad}\mathcal{E}(X_k) \qquad \triangleright \text{ See Algorithm 1}$$

3: Compute noise term

$$B = \sqrt{\tfrac{2\Delta t}{\beta}} \begin{bmatrix} B_{11} & B_{12} \\ B_{12}^T & 0 \end{bmatrix} + \tfrac{\Delta t}{\beta} \sum_{i=1}^{p} \tfrac{1}{\lambda_i} \begin{bmatrix} 0 & 0 \\ 0 & I_{n-p} \end{bmatrix}$$

4: Obtain the new iterate by retraction

$$X_{k+1} = \mathrm{P}_{\mathcal{S}_+^{n,p}}(X_k - \Delta t \xi_k + Q_k B Q_k^T) \qquad \triangleright \text{ See Algorithm 2}$$

5: **end for**

Remark 26. The mean curvature correction term is necessary for avoiding rank deficient samples in the following sense. A sampling scheme on $\mathcal{S}_+^{n,p}$ might generate a sample X with a rank numerically close to $p-1$, and the mean curvature correction term in the scheme (4.1.7) would be huge if $\lambda_p \to 0$, thus it will force iterate X_k to stay away from the boundary of $\mathcal{S}_+^{n,p}$.

Remark 27. Notice that the complexity of computing SVD of $X + Z$ in Algorithm 2 would be $\mathcal{O}(n^3)$ in a naive implementation. For a Riemannian gradient method, if $Z \in T_{X_k}\mathcal{S}_+^{n,p}$, a compact implementation of computing $P_{\mathcal{S}_+^{n,p}}(X + Z)$ in [96] is only $\mathcal{O}(np^2) + \mathcal{O}(p^3)$, which is no longer possible for the Langevin Monte Carlo scheme (4.1.7) due to the mean curvature correction term in the normal space. On the other hand, if Lanczos type algorithm is used for computing to top p eigen-componenes of $X + Z$, it seems possible to explore the special structure in (4.3.1) to find a more efficient implementation, but we do not consider a more compact implementation in this paper.

4.3.2 Scheme BW for the Bures-Wasserstein metric

The Riemannian gradient and a simple retraction operator

Given a smooth energy function $\mathcal{E}(X)$ defined on $\mathcal{S}_+^{n,p}$, the corresponding function h on $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$ satisfies

$$\begin{aligned} h : \mathbb{R}_*^{n\times p}/\mathcal{O}_p &\to \mathbb{R} \\ \pi(Y) &\mapsto \mathcal{E}(\tilde{\beta}(\pi(Y))) = \mathcal{E}(\beta(Y)) = \mathcal{E}(YY^T). \end{aligned} \tag{4.3.3}$$

Observe that the function $F(Y) := \mathcal{E}(YY^T)$ satisfies $F(Y) = h \circ \pi(Y) = \mathcal{E} \circ \beta(Y)$. The Riemannian gradient of h at $\pi(Y)$ is a tangent vector in $T_{\pi(Y)}\mathbb{R}_*^{n\times p}/\mathcal{O}_p$. The next theorem is given in [1, Section 3.6.2], showing that the horizontal lift of $\mathrm{grad}h(\pi(Y))$ can be obtained from the Riemannian gradient of F defined on $\mathbb{R}_*^{n\times p}$.

Theorem 18. *The horizontal lift of the gradient of h at $\pi(Y)$ is the Riemannian gradient of F at Y. That is,*

$$\overline{\mathrm{grad}h(\pi(Y))}_Y = \mathrm{grad}F(Y).$$

For the Bures-Wasserstein metric, the following result is proven in [96]:

Proposition 28. Let $\mathcal{E}$ be a smooth real-valued function defined on $\mathcal{S}_+^{n,p}$ and let $F : \mathbb{R}_*^{n\times p} \to \mathbb{R} : Y \mapsto \mathcal{E}(YY^T)$. Assume $YY^T = X$. Then the Riemannian gradient of F is given by

$$\mathrm{grad}F(Y) = 2\nabla\mathcal{E}(YY^T)Y$$

where $\nabla\mathcal{E}(\cdot)$ is the gradient of $\mathcal{E}$ w.r.t. X.

In [69, Prop. A.8], the relationship between the horizontal lifts of the quotient tangent vector $\xi_{\pi(Y)}$ lifted at different representatives in $[Y]$ is given:

Lemma 19. *Let η be a vector field on $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$, and let $\bar{\eta}$ be the horizontal lift of η. Then for each $Y \in \mathbb{R}_*^{n\times p}$, we have*

$$\bar{\eta}_{YO} = \bar{\eta}_Y O$$

for all $O \in \mathcal{O}_p$.

The retraction on the quotient manifold $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$ can be defined using the retraction on the total space $\mathbb{R}_*^{n\times p}$. For any $A \in T_Y\mathbb{R}_*^{n\times p}$ and a step size $\tau > 0$,

$$\overline{R}_Y(\tau A) := Y + \tau A,$$

is a retraction on $\mathbb{R}_*^{n\times p}$ if $Y + \tau A$ remains full rank, which is ensured for small enough τ. Then Lemma 19 indicates that $\overline{R}$ satisfies the conditions of [1, Prop. 4.1.3], which implies that

$$R_{\pi(Y)}(\tau\eta_{\pi(Y)}) := \pi(\overline{R}_Y(\tau\overline{\eta}_Y)) = \pi(Y + \tau\overline{\eta}_Y) \tag{4.3.4}$$

defines a retraction on the quotient manifold $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$ for a small enough step size $\tau > 0$.

Finally, we give an example of what these results imply by considering the Riemannian gradient descent method for minimizing $\mathcal{E}(X)$ over $(\mathcal{S}_+^{n,p}, g_{BW})$. With the simple retraction (4.3.4), the Riemannian gradient descent method for minimizing the function $h[\pi(Y)]$ on $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$ is given by

$$Y_{k+1} = Y_k - \Delta t 2\nabla\mathcal{E}(Y_k Y_k^T)Y_k,$$

which is the simple Burer-Monteiro gradient descent method for minimizing $\mathcal{E}(X)$ over $\mathcal{S}_+^{n,p}$. See Section 5.1 in [96] for details.

A simple Riemannian Langevin Monte Carlo scheme

With the Euler-Maruyama discretization for SDE (4.2.6), and the simple retraction and Riemannian gradient as given previously, a simple Riemannian Langevin Monte Carlo scheme for approximating the Riemannian SDE (4.2.6) on the Riemannian manifold $(S_+^{n,p}, g_{BW})$ can be given as

$$Y_{k+1} = Y_k - \Delta t 2\nabla\mathcal{E}(Y_k Y_k^T)Y_k + \sqrt{\frac{2\Delta t}{\beta}}B_k + \frac{\Delta t}{\beta}U\left[\sum_{j:j\neq i}\frac{\sigma_i}{\sigma_i^2+\sigma_j^2}\right]_{ii}V^T, \tag{4.3.5}$$

where B_k is n-by-p matrix with i.i.d. $\mathcal{N}(0,1)$ entries and $Y_k = U\Sigma V^T$ is the compact SVD of Y with singular values $\sigma_i > 0$ for $i = 1, 2, \cdots, p$.

Notice that all operations are performed in the space of size $n \times p$. For finding compact SVD of Y, one can first compute QR decomposition of Y, which costs $\mathcal{O}(np^2) + \mathcal{O}(p^3)$. Then compute SVD of size $p \times p$, which is $\mathcal{O}(p^3)$. So the complexity of this scheme is $\mathcal{O}(np^2) + \mathcal{O}(p^3)$ for each iteration. For large n and small p, Scheme BW should be cheaper than Scheme E in each iteration, but they generate different samples for different Gibbs distributions which depend on the metric, i.e., Scheme BW cannot replace Scheme E for generating Gibbs distribution defined by embedded geometry.

4.4 Examples with analytical formulae

In this section, we provide a few examples with analytical formulae so that they can be used in numerical experiments for testing the two schemes (4.3.1) and (4.3.5) on the Gibbs distribution.

For the rest of this section, $X = Q\Lambda Q^T \in \mathcal{S}_+^{n,p}$ denotes the full SVD with descending eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$.

4.4.1 Scalar random variables

Let X be a random variable satisfying the Gibbs distribution on $\mathcal{S}_+^{n,p}$ with dimension $N = np - \frac{p(p-1)}{2}$ under either metric, then X is a matrix-valued random variable. For convenience, we consider a scalar random variable $D = D(X)$ as a function of $X \in \mathcal{S}_+^{n,p}$, e.g., $D = \|X\|_F$ where $\|\cdot\|_F$ is the matrix Frobenius norm.

We consider the distribution function for the scalar random variable D:

$$\Pr[D < d] = \frac{1}{Z_\beta} \int\limits_{U_d} \mathrm{e}^{-\beta \mathcal{E}} \mathrm{d}V, \quad Z_\beta = \int\limits_{\mathcal{M}} \mathrm{e}^{-\beta \mathcal{E}} \mathrm{d}V, \tag{4.4.1}$$

where $U_d := \{X \in \mathcal{S}_+^{n,p} | D(X) < d\}$ is the domain of integral. For simplicity we only consider symmetric functions such that the random variable D, the energy function $\mathcal{E}$, and the volume form are all invariant under the group action by the orthogonal group $\mathcal{O}_n$. We consider an energy function $\mathcal{E}$ satisfying $\mathcal{E}(X) = \mathcal{E}(OXO^T)$, $\forall\, O \in \mathcal{O}_n$, so that Gibbs distribution function only depends on the spectrum of X when considering (4.4.1) with $D = ||X||_F = \sqrt{\lambda_1^2 + \cdots + \lambda_p^2}$. Since $\mathcal{O}_n$ is an isometry group for both metrics g_E and g_{BW}, the volume form $\mathrm{d}V$ in the two cases is also invariant under $\mathcal{O}_n$ action.

Notice that Q and Λ can be used as coordinates of the manifold $\mathcal{S}_+^{n,p}$. The volume form expressed by coordinates Q and Λ is given by

$$\mathrm{d}V = \sqrt{\det g}\Big(\prod_{i=1}^{p} \mathrm{d}\lambda_i\Big)\mathrm{d}\mu_{\mathcal{O}_n}, \tag{4.4.2}$$

where $\mu_{\mathcal{O}_n}$ is the Haar measure on $\mathcal{O}_n$, and g is the matrix of metric g_E or g_{BW} expressed under coordinate Q and λ. For g_E its determinant $\det g$ is

$$\det g = \Big(\prod_{1\le i<j\le p} |\lambda_i - \lambda_j|^2\Big)\Big(\prod_{1\le i\le p} \lambda_i^{2(n-p)}\Big), \tag{4.4.3}$$

and for g_{BW} it is

$$\det g = \Big(\prod_{1\le i<j\le p} \frac{|\lambda_i - \lambda_j|^2}{\lambda_i + \lambda_j}\Big)\Big(\prod_{1\le i\le p} \lambda_i^{(n-p)}\Big). \tag{4.4.4}$$

So for g_E the distribution $\Pr[D < d]$ is expressed as

$$\begin{aligned}\Pr[D < d] =& \frac{1}{Z_\beta} \int_{||X||_F<d} \mathrm{e}^{-\beta\mathcal{E}}\mathrm{d}V \\ \propto& \int_{\substack{\sum_{i=1}^{p} \lambda_i^2 < d^2 \\ \lambda_i>0, i=1,\ldots,p}} \mathrm{e}^{-\beta\mathcal{E}(\lambda_1,\ldots,\lambda_p)}\Big(\prod_{1\le i<j\le p} |\lambda_i - \lambda_j|\Big)\Big(\prod_{1\le i\le p} \lambda_i^{n-p}\Big)\mathrm{d}\lambda_1\cdots\mathrm{d}\lambda_p,\end{aligned} \tag{4.4.5}$$

where we have used the fact that the integrand does not depend on the coordinate $Q \in \mathcal{O}_n$, so the integral of $\mu_{\mathcal{O}_n}$ only provides a constant coefficient. As we could always renormalize $\Pr[D < d]$ by considering the quotient $\frac{\Pr[D<d]}{\Pr[D<\infty]}$, we only need the dependence of the integral on parameter d.

Similarly, for the Bures-Wasserstein metric g_{BW} we have

$$\Pr[D < d] \propto \int\limits_{\substack{\sum_{i=1}^{p} \lambda_i^2 < d^2 \\ \lambda_i > 0, i=1,\dots,p}} \mathrm{e}^{-\beta \mathcal{E}(\lambda_1,\dots,\lambda_p)} \Big(\prod_{1 \le i < j \le p} \frac{|\lambda_i - \lambda_j|}{\sqrt{\lambda_i + \lambda_j}} \Big) \Big(\prod_{1 \le i \le p} \lambda_i^{\frac{n-p}{2}} \Big) \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_p \tag{4.4.6}$$

Next we give a few energy functions.

4.4.2 Example I: $\mathcal{E}(X) = \frac{1}{2}||X||_F^2$

This is the simplest example. Using the general expression above, for embedded geometry g_E we have

$$\begin{aligned}
\Pr[D < d] \propto & \int\limits_{\substack{\sum_{i=1}^{p} \lambda_i^2 < d^2 \\ \lambda_i > 0, i=1,\dots,p}} \mathrm{e}^{-\frac{\beta}{2} \sum_{i=1}^{p} \lambda_i^2} \Big(\prod_{1 \le i < j \le p} |\lambda_i - \lambda_j| \Big) \Big(\prod_{1 \le i \le p} \lambda_i^{n-p} \Big) \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_p \\
= & \int_0^d \mathrm{e}^{-\frac{\beta}{2}\rho^2} \rho^{N-1} \Big(\int\limits_{S_+^{p-1}} \prod_{1 \le i < j \le p} |\omega_i - \omega_j| \prod_{i=1}^{p} |\omega_i|^{n-p} \prod_{i=1}^{p} \mathrm{d}\omega \Big) \mathrm{d}\rho \\
= & \Big(\int\limits_{S_+^{p-1}} \prod_{1 \le i < j \le p} |\omega_i - \omega_j| \prod_{i=1}^{p} |\omega_i|^{n-p} \prod_{i=1}^{p} \mathrm{d}\omega \Big) \int_0^d \mathrm{e}^{-\frac{\beta}{2}\rho^2} \rho^{N-1} \mathrm{d}\rho \\
\propto & \int_0^d \mathrm{e}^{-\frac{\beta}{2}\rho^2} \rho^{N-1} \mathrm{d}\rho,
\end{aligned} \tag{4.4.7}$$

where we have used the spherical coordinate for $(\lambda_1, \dots, \lambda_p) = \rho\omega$, with $\rho = \sqrt{\sum_{i=1}^{p} \lambda_i^2}$ being the radius and $\omega \in S_+^{p-1} = S^{p-1} \cap \mathbb{R}_+^p$ being the coordinate on the positive orthant of unit sphere.

For g_{BW}, similarly we have

$$\Pr[D < d] \propto \int_0^d \mathrm{e}^{-\beta\rho^2} \rho^{\frac{N}{2}-1} \mathrm{d}\rho. \tag{4.4.8}$$

Now we can see that $\beta D^2 = \beta||X||_F^2$ is subject to $\chi^2(N)$ distribution for the embedded metric g_E, and $\chi^2(\frac{N}{2})$ distribution for the Bures-Wasserstein metric.

4.4.3 Example II: $\mathcal{E}(X) = \mathrm{Tr}(X \log X)$

We consider the von Neumann entropy

$$\mathcal{E}(X) = \mathrm{Tr}(X \log X) = \sum_{i=1}^{p} \lambda_i \log \lambda_i$$

and construct a more interesting example. The minimizers of $\mathcal{E}(X) = \mathrm{Tr}(X \log X)$ on $\mathcal{S}_+^{n,p}$ are matrices $X \in \mathcal{S}_+^{n,p}$ with spectrum $\lambda_1 = \cdots = \lambda_p = \mathrm{e}^{-1}$.

The random variable we consider is still $D = ||X||_F$. Since $\mathcal{E}(X) = \mathrm{Tr}(X \log X) = \sum\limits_{i=1}^{p} \lambda_i \log \lambda_i$ only depends on spectrum, the argument in the previous section about integral on $\mathcal{O}_n$ still applies. Similar to (4.4.7), for g_E we have

$$\begin{aligned}
\Pr(D < d) &= \int\limits_{\substack{\sum_{i=1}^{p} \lambda_i^2 < d^2 \\ \lambda_i > 0, i=1,\ldots,p}} \mathrm{e}^{-\beta \sum_{i=1}^{p} \lambda_i \log \lambda_i} \prod_{1 \le i < j \le p} |\lambda_i - \lambda_j| \prod_{i=1}^{p} |\lambda_i|^{n-p} \prod_{i=1}^{p} \mathrm{d}\lambda_i \\
&= \int\limits_{\substack{\sum_{i=1}^{p} \lambda_i^2 < d^2 \\ \lambda_i > 0, i=1,\ldots,p}} \prod_{1 \le i < j \le p} |\lambda_i - \lambda_j| \prod_{i=1}^{p} |\lambda_i|^{n-p-\beta\lambda_i} \prod_{i=1}^{p} \mathrm{d}\lambda_i,
\end{aligned}$$

and for g_{BW} we have

$$\Pr(D<d)=\int\limits_{\substack{\sum_{i=1}^{p}\lambda_i^2<d^2\\ \lambda_i>0,i=1,\ldots,p}}\prod_{1\le i<j\le p}\frac{|\lambda_i-\lambda_j|}{\sqrt{\lambda_i+\lambda_j}}\prod_{i=1}^{p}|\lambda_i|^{\frac{n-p}{2}-\beta\lambda_i}\prod_{i=1}^{p}\mathrm{d}\lambda_i.$$

Although we do not have a closed expression for both cases, such integrals can be easily approximated by an accurate quadrature.

4.4.4 Example III: $\mathcal{E}(X)=\frac{1}{2}||X-A||_F^2$

We consider a quadratic function $\mathcal{E}(X)=\frac{1}{2}||X-A||_F^2$ where $A\in\mathcal{S}_+^{n,p}$ with $D=||X-A||_F$. In this example, $\mathcal{O}_n$ symmetry does not hold, and we can only make an estimate of the distribution function.

The random variable D we are considering now is $D=||X-A||_F$, its distribution function is evaluated as

$$\Pr(D<d)\propto\int_{U_d}\mathrm{e}^{-\frac{\beta}{2}D^2}\mathrm{d}V \tag{4.4.9}$$

where $U_d=\{X\in\mathcal{S}_+^{n,p}|D(X)<d\}$. Using delta function, formally we can simplify

the integral to

$$
\begin{aligned}
\Pr(D < d) &\propto \int_{\mathcal{M}} \mathbf{1}_{\{D<d\}} \mathrm{e}^{-\frac{\beta}{2}D^2} \mathrm{d}V && (4.4.10) \\
&= \int_{\mathcal{M}} (\int_0^\infty \mathbf{1}_{\{\rho<d\}} \mathrm{e}^{-\frac{\beta}{2}\rho^2} \delta(D-\rho)\mathrm{d}\rho)\mathrm{d}V \\
&= \int_0^\infty \mathbf{1}_{\{\rho<d\}} \mathrm{e}^{-\frac{\beta}{2}\rho^2} (\int_{\mathcal{M}} \delta(D-\rho)\mathrm{d}V)\mathrm{d}\rho \\
&= \int_0^d \mathrm{e}^{-\frac{\beta}{2}\rho^2} (\int_{\mathcal{M}} \frac{\mathrm{d}}{\mathrm{d}\rho} \mathbf{1}_{\{D-\rho\}} \mathrm{d}V)\mathrm{d}\rho \\
&= \int_0^d \mathrm{e}^{-\frac{\beta}{2}\rho^2} \frac{\mathrm{d}}{\mathrm{d}\rho} (\int_{\mathcal{M}} \mathbf{1}_{\{D-\rho\}} \mathrm{d}V)\mathrm{d}\rho \\
&= \int_0^d \mathrm{e}^{-\frac{\beta}{2}\rho^2} \frac{\mathrm{d}}{\mathrm{d}\rho} V_D(\rho)\mathrm{d}\rho
\end{aligned}
$$

(4.4.11)

where $V_D(\rho) = \int_{\mathcal{M}} \mathbf{1}_{\{D<\rho\}} \mathrm{d}V = \int_{D<\rho} \mathrm{d}V$.

In general it is difficult to calculate $\int_{D<\rho} \mathrm{d}V$, but we consider the following approximation. Consider the volume of the ball $B_A^{n,p}(r) = B_A(r) \cap \mathcal{S}_+^{n,p}$, where

$$B_A(r) = \big\{X \in \mathcal{S}^{n \times n} : ||X - A||_F < r\big\}.$$

It is difficult to compute $\mathrm{Vol}(B_A^{n,p}(r))$, but we propose the following estimate, for fixed $A \in \mathcal{S}_+^{n,p}$:

$$\mathrm{Vol}(B_{cA}^{n,p}(r)) \approx \alpha r^N, \quad c \gg 1, \tag{4.4.12}$$

where α is a constant that does not depend on r, N is the dimension of $\mathcal{S}_+^{n,p}$. For g_E, α is exactly the volume of unit ball in $\mathbb{R}^N$, while for g_{BW}, α depends on dimension N and $A \in \mathcal{S}_+^{n,p}$.

For the embedded geometry, the approximation (4.4.12) can be justified by the following arguments:

1. The second fundamental form $\boldsymbol{I\!I}_{cA}$ of the manifold is vanishing for fixed A and $c \to \infty$. See [93].

2. The Riemannian curvature tensor of ambient space $\mathcal{S}^{n\times n}$ is 0. Applying the Gauss equation [30, Prop 3.1] we can express the Riemannian curvature tensor R of $(\mathcal{S}_+^{n,p}, g_E)$ in terms of its second fundamental form $\boldsymbol{I\!I}$:

$$\langle R(\boldsymbol{x},\boldsymbol{y})\boldsymbol{z},\boldsymbol{w}\rangle = -\langle \boldsymbol{I\!I}(\boldsymbol{x},\boldsymbol{z}), \boldsymbol{I\!I}(\boldsymbol{y},\boldsymbol{w})\rangle + \langle \boldsymbol{I\!I}(\boldsymbol{x},\boldsymbol{w}), \boldsymbol{I\!I}(\boldsymbol{y},\boldsymbol{z})\rangle, \quad (4.4.13)$$
$$\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\boldsymbol{w} \in T\mathcal{S}_+^{n,p}, \qquad \langle U,V\rangle = \mathrm{Tr}(UV^T) \text{ is the metric in } \mathcal{S}^{n\times n},.$$

   Thus, with vanishing $\boldsymbol{I\!I}$ we have vanishing Riemannian curvature tensor, and zero sectional curvature.

3. Vanishing extrinsic curvature and intrinsic curvature means that the neighborhood is approximately an Euclidean space, so the ball $B_{cA}^{n,p}(r)$ is approximately just a ball in $\mathbb{R}^N$ and has volume αr^N, with α being the volume of a unit ball.

For the g_{BW} metric, following similar arugments, we can get the same approximation (4.4.12). We emphasize that the approximation (4.4.12) is accurate only if c is large enough. Putting all this together, when A has eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p \gg 1$, we have the following

$$\Pr(D<d) \propto \int\limits_{D<d} \mathrm{e}^{-\frac{\beta}{2}D^2}\mathrm{d}V = \int_0^d \mathrm{e}^{-\frac{\beta}{2}\rho^2}\frac{\mathrm{d}}{\mathrm{d}\rho}\Big(\int\limits_{D<\rho}\mathrm{d}V\Big)\mathrm{d}\rho \mathrel{\overset{\propto}{\sim}} \int_0^t \mathrm{e}^{-\frac{\beta}{2}\rho^2}\rho^{N-1}\mathrm{d}\rho, \quad (4.4.14)$$

where $\overset{\propto}{\sim}$ stands for *being approximately proportional to*.

4.4.5 MCMC numerical integration

It is well known that MCMC can be used for integrating a function numerically, and that one of the main advantages is that the convergence rate is independent of the dimension. Both schemes in this paper are MCMC type sampling schemes on the manifold. Suppose we have generated samples X_i satisfying the Gibbs distribution on the manifold, e.g.,

$$X_i \sim \frac{1}{Z_\beta} e^{-\beta\mathcal{E}(X)} \mathrm{d}V_g,$$

where $Z_\beta = \int\limits_{S_+^{n,p}} \mathrm{e}^{-\beta\mathcal{E}(X)}\mathrm{d}V$ is an unknown normalization factor and $\mathrm{d}V$ is the volume form depending on the metric. Then for approximating the integral of a nice function $f(X)$ on the same manifold $\int\limits_{S_+^{n,p}} f(X)\mathrm{d}V$, we can use

$$\frac{1}{m}\sum_{i=1}^{m} f(X_i)\mathrm{e}^{\beta\mathcal{E}(X_i)} \approx \frac{\int\limits_{S_+^{n,p}} f(X)\mathrm{d}V}{\int\limits_{S_+^{n,p}} \mathrm{e}^{-\beta\mathcal{E}(X)}\mathrm{d}V} = \frac{1}{Z_\beta}\int\limits_{S_+^{n,p}} f(X)\mathrm{d}V, \tag{4.4.15}$$

because each $f(X_i)\mathrm{e}^{\beta\mathcal{E}(X_i)}$ is a random variable with expectation

$$\mathbb{E}\left[f(X_i)\mathrm{e}^{\beta\mathcal{E}(X_i)}\right] = \frac{1}{Z_\beta}\int\limits_{S_+^{n,p}} f(X_i)\mathrm{e}^{\beta\mathcal{E}(X_i)}\mathrm{e}^{-\beta\mathcal{E}(X_i)}\mathrm{d}V,$$

and the left hand side is a random variable with expectation

$$\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} f(X_i)\mathrm{e}^{\beta E(X_i)}\right] = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[f(X_i)\mathrm{e}^{\beta E(X_i)}\right] = \frac{1}{Z_\beta}\int\limits_{S_+^{n,p}} f(X)\mathrm{d}V,$$

where the expectation $\mathbb{E}[\cdot]$ is taken w.r.t. Gibbs distribution under corresponding metric.

So using the generated samples X_i, we can approximate the integral $\int\limits_{\mathcal{S}_+^{n,p}} f(X)\mathrm{d}V$

up to a constant Z_β that does not depend on $f(X)$. Notice that the additional advantage of Monte Carlo type quadrature on a manifold is that we do not need to know what $\mathrm{d}V$ is. On the other hand, Z_β cannot be approximated by the same approach. Though we do not consider any specific application for numerical integration, equation (4.4.15) can be used as one way to validate the Riemannian Langevin Monte Carlo schemes.

For the following special functions, it is possible to calculate exact integrals. For the energy function $\mathcal{E}(X) = \frac{1}{2}\|X\|_F^2$, and a special integrand $f(X) = \|X\|_F^k \mathrm{e}^{-\frac{\alpha}{m}\|X\|_F^m}$ with $k > -N, m > 2, \alpha > 0$, using the results in 4.4.2, the distribution of $D = ||X||_F$ is

$$\text{for metric } g_E: \quad \Pr[D < d] \propto \int_0^d \mathrm{e}^{-\frac{\beta}{2}\rho^2}\rho^{N-1}\mathrm{d}\rho, \tag{4.4.16}$$

$$\text{for metric } g_{BW}: \quad \Pr[D < d] \propto \int_0^d \mathrm{e}^{-\frac{\beta}{2}\rho^2}\rho^{\frac{N}{2}-1}\mathrm{d}\rho, \tag{4.4.17}$$

so the integral on the manifold could be expressed by expectation of a random variable, which leads to

$$\begin{aligned}
\text{for } g_E:& \frac{1}{Z_\beta}\int_{\mathcal{S}_+^{n,p}} f(X)\mathrm{d}V = \mathbb{E}[f(X)\mathrm{e}^{\frac{\beta}{2}||X||_F^2}] = \mathbb{E}[D^k\mathrm{e}^{-\frac{\alpha}{m}D^m}\mathrm{e}^{\frac{\beta}{2}D^2}] \\
&= \frac{\int_0^\infty \rho^k \mathrm{e}^{-\frac{\alpha}{m}\rho^m+\frac{\beta}{2}\rho^2}\rho^{N-1}\mathrm{e}^{-\frac{\beta}{2}\rho^2}\mathrm{d}\rho}{\int_0^\infty \rho^{N-1}\mathrm{e}^{-\frac{\beta}{2}\rho^2}\mathrm{d}\rho} = \frac{\frac{1}{m}(\alpha/m)^{-\frac{k+N}{m}}\Gamma((k+N)/m)}{\frac{1}{2}(\beta/2)^{-N/2}\Gamma(N/2)} && (4.4.18)\\
\text{for } g_{BW}:& \frac{1}{Z_\beta}\int_{\mathcal{S}_+^{n,p}} f(X)\mathrm{d}V = \mathbb{E}[f(X)\mathrm{e}^{\frac{\beta}{2}||X||_F^2}] = \mathbb{E}[D^k\mathrm{e}^{-\frac{\alpha}{m}D^m}\mathrm{e}^{\frac{\beta}{2}D^2}] \\
&= \frac{\int_0^\infty \rho^k \mathrm{e}^{-\frac{\alpha}{m}\rho^m+\frac{\beta}{2}\rho^2}\rho^{\frac{N}{2}-1}\mathrm{e}^{-\frac{\beta}{2}\rho^2}\mathrm{d}\rho}{\int_0^\infty \rho^{\frac{N}{2}-1}\mathrm{e}^{-\frac{\beta}{2}\rho^2}\mathrm{d}\rho} = \frac{\frac{1}{m}(\alpha/m)^{-\frac{k+N/2}{m}}\Gamma((k+N/2)/m)}{\frac{1}{2}(\beta/2)^{-N/4}\Gamma(N/4)}.
\end{aligned}$$

(4.4.19)

4.5 Numerical tests

In this section we test the samples generated by the two Riemannian Langevin Monte Carlo schemes (4.3.1) and (4.3.5) on the examples constructed in the previous section. The samples are generated by the following procedure: we run the iterative schemes (4.3.1) or (4.3.5) for sufficiently many $\tilde{m}$ iterations then take the last m iterates as the samples for the Gibbs distribution. Both $\tilde{m}$ and m should be chosen such that the $(\tilde{m}-m)$-th iterate has already reached equilibrium e.g., $\tilde{m}$ is $6,000,000$ and m is $5,000,000$ for specially chosen energy functions and parameters β.

Now suppose we have generated samples $X_i \in \mathcal{S}_+^{n,p}$ $(i = 1, \cdots, m)$ for either metric. In order to test or show the numerical convergence to the Gibbs distribution, we will consider two kinds of numerical tests.

The first kind of tests is to test on the scalar random variable $D(X) = \|X\|_F$ or $D(X) = \|X - A\|_F$ as described in Section 4.4. Then we compare the cumulative distribution function (CDF) of the random variable D with its empirical CDF calculated from the MCMC samples.

Denote the true CDF of D by $F_D(t) := \Pr(D \le t)$. The empirical CDF of samples is

$$\hat{F}_D(t) := \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{D(X_i) \le t},$$

where $\mathbb{1}_{D(X_i)\le t}$ takes value 1 if $D(X_i) \le t$, and value 0 if otherwise. The Kolmogorov–Smirnov test statistic (K-S statistic) is defined by

$$KS_D := \sup_t \left| F_D(t) - \hat{F}_D(t) \right|. \tag{4.5.1}$$

In our numerical tests, we compute the KS statistic by taking the maximum difference of F_D and $\hat{F}_D$ at 100 equally spaced points in the interval $[0, t_{max}]$ where

$F_D(t_{max}) \approx 1$.

The second kind of tests is on the integral examples in Section 4.4.5, let X be a random variable satisfying Gibbs distribution on the manifold $\mathcal{S}_+^{n,p}$ under either metric. Define

$$\mu := \mathbb{E} f(X)\mathrm{e}^{\beta\mathcal{E}(X)} = \frac{1}{Z_\beta}\int\limits_{\mathcal{S}_+^{n,p}} f(X)\mathrm{d}V.$$

Given m samples $X_i \in \mathcal{S}_+^{n,p}$, we define

$$\hat{\mu}_m := \frac{1}{m}\sum_{i=1}^{m} f(X_i)\mathrm{e}^{\beta\mathcal{E}(X_i)}. \tag{4.5.2}$$

Notice that samples generated by MCMC are not independent. If we assume

$$\sigma^2 := \mathrm{var}\left(f(X_1)\mathrm{e}^{\beta\mathcal{E}(X_1)}\right) + 2\sum_{k=1}^{\infty}\mathrm{cov}\left(f(X_1)\mathrm{e}^{\beta\mathcal{E}(X_1)}, f(X_{1+k})\mathrm{e}^{\beta\mathcal{E}(X_{1+k})}\right) < \infty,$$

then by the Markov Chain Central Limit Theorem[51, 37], as $m \to \infty$, we have

$$\sqrt{m}(\hat{\mu}_m - \mu) \to \mathcal{N}(0, \sigma^2) \tag{4.5.3}$$

where the convergence is in the sense of distribution. Thus if $m \gg 1$, $\frac{\hat{\mu}_m - \mu}{\mu}$ roughly follows the distribution $\mathcal{N}(0, \mathcal{O}(\frac{1}{m}))$ and the relative error term $\left|\frac{\hat{\mu}_m - \mu}{\mu}\right|$ roughly follows the folded normal distribution with mean $\mathcal{O}(\frac{1}{\sqrt{m}})$ and variance $\mathcal{O}(\frac{1}{m})$. Hence we can use $\hat{\mu}_m$ defined in (4.5.2) to estimate $\mu = \frac{1}{Z_\beta}\int\limits_{\mathcal{S}_+^{n,p}} f(X)\mathrm{d}V$, and the relative error is $\mathcal{O}(\frac{1}{\sqrt{m}})$.

4.5.1 Numerical validation of the scalar variable $D(X)$

The manifold $\mathcal{S}_+^{n,p}$ has dimension $N = np - p(p-1)/2$. For both metrics, we consider three examples in Section 4.4 with special energy functions $\mathcal{E}$ in the Gibbs distribution $\mathrm{e}^{-\beta\mathcal{E}}$ and the CDF for the scalar variable $D(X)$:

1. Example I: $\mathcal{E}(X) = \frac{1}{2}\|X\|_F^2$ with the CDF for $D(X) = \|X\|_F$:

$$\text{For } g_E: \quad F_D(t) = \Pr(\|X\|_F \le t) \propto \int_0^t \mathrm{e}^{-\frac{\beta}{2}\rho^2}\rho^{N-1}\mathrm{d}\rho,$$

$$\text{For } g_{BW}: \quad F_D(t) = \Pr(\|X\|_F \le t) \propto \int_0^t \mathrm{e}^{-\frac{\beta}{2}\rho^2}\rho^{N/2-1}\mathrm{d}\rho.$$

2. Example II: $\mathcal{E}(X) = \mathrm{Tr}(X\log X)$ with the CDF $F_D(t) = \Pr(\|X\|_F \le t)$ for $D(X) = \|X\|_F$:

$$\text{For } g_E: \quad F_D(t) \propto \int\limits_{\substack{\sum_{i=1}^{p}\lambda_i^2 < t^2 \\ \lambda_i > 0, i=1,\ldots,p}} \prod_{1\le i<j\le p}|\lambda_i - \lambda_j| \prod_{i=1}^{p}|\lambda_i|^{n-p-\beta\lambda_i}\prod_{i=1}^{p}\mathrm{d}\lambda_i,$$

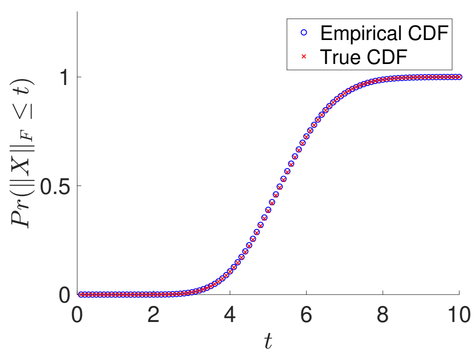
$$\text{For } g_{BW}: \quad F_D(t) \propto \int\limits_{\substack{\sum_{i=1}^{p}\lambda_i^2 < t^2 \\ \lambda_i > 0, i=1,\ldots,p}} \prod_{1\le i<j\le p}\frac{|\lambda_i - \lambda_j|}{\sqrt{\lambda_i + \lambda_j}} \prod_{i=1}^{p}|\lambda_i|^{\frac{n-p-1}{2}-\beta\lambda_i}\prod_{i=1}^{p}\mathrm{d}\lambda_i.$$

   which is a p-fold integral and can be approximated accurately by quadrature such as Simpson's rule for relatively small values of p, e.g., $p = 2, 3$.

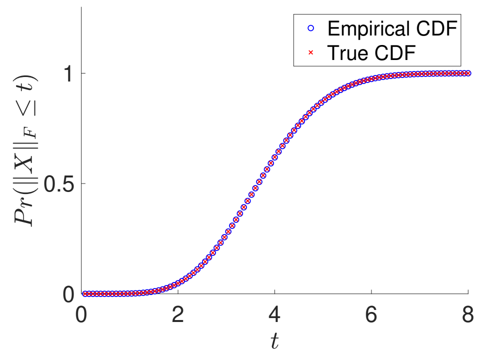
3. Example III: $\mathcal{E}(X) = \frac{1}{2}\|X - A\|_F^2$ where $A \in \mathcal{S}_+^{n,p}$ has eigenvalues $\lambda_1 \ge \cdots \ge \lambda_p \gg 1$, with the CDF for $D(X) = \|X - A\|_F$:

$$\text{For both } g_E \text{ and } g_{BW}: \quad F_D(t) = \Pr(\|X - A\|_F \le t) \mathrel{\underset{\sim}{\propto}} \int_0^t \mathrm{e}^{-\frac{\beta}{2}\rho^2}\rho^{N-1}\mathrm{d}\rho.$$

In implementation of the scheme, the step size Δt and β in schemes (4.3.1) and (4.3.5) are two parameters that need to be tuned to reach equilibrium with reasonable computing time. We first use a numerically stable Δt then adjust β so that the noise term has reasonable variance. And of course one needs a sufficient large number of iterations for schemes (4.3.1) and (4.3.5) to reach their equilibrium state, and a sufficient large number m of samples to observe numerical convergence toward the Gibbs distribution through the scalar random variable D, e.g., the KS statistic (4.5.1) should be small. See Figure 4.5.1, Figure 4.5.2, Figure 4.5.3, and Figure 4.5.4 for the numerical results.



(a) Scheme E (4.3.1) on $(\mathcal{S}_+^{n,p}, g_E)$ with $\Delta t = 0.001$ and $\beta = 0.4$. The error between two CDFs is $KS = 0.0054$.



(b) Scheme BW (4.3.5) on $(\mathcal{S}_+^{n,p}, g_{BW})$ with $\Delta t = 0.001$ and $\beta = 0.4$. The error between two CDFs is $KS = 0.0023$.

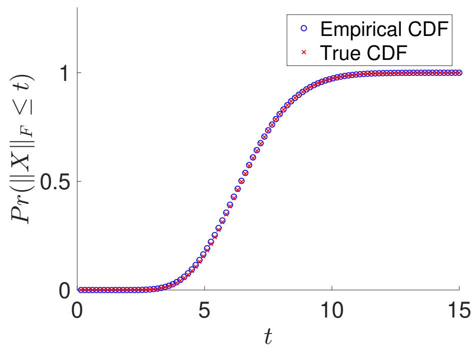
Figure 4.5.1: Example I: $\mathcal{E}(X) = \frac{1}{2}\|X\|_F^2, n = 5, p = 3$ and manifold dimension is $N = 12$. The empirical CDF is computed by $5E6$ MCMC samples generated after $6E6$ iterations of the Riemannian Langevin Monte Carlo schemes. Both CDFs of scheme E and scheme BW are evaluated at 100 equally spaced points on $[0, 10]$ and $[0, 8]$, respectively, and the difference can be measured by the KS statistic (4.5.1).

4.5.2 MCMC numerical integration

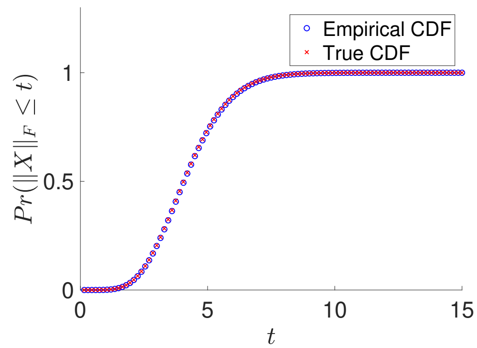
We consider special cases $k = 0, m = 2$ in the examples (4.4.18) and (4.4.19), then (4.4.18) reduces to $(\frac{\beta}{\alpha})^{N/2}$ and (4.4.19) reduces to $(\frac{\beta}{\alpha})^{N/4}$. In other words, we may verify the numerical convergence of samples X_i to Gibbs distribution by verifying

$$\text{For } g_E: \quad \frac{1}{m}\sum_{i=1}^{m} \mathrm{e}^{-\frac{\alpha-\beta}{2}\|X_i\|_F^2} \to (\frac{\beta}{\alpha})^{N/2}, \tag{4.5.4}$$

$$\text{For } g_{BW}: \quad \frac{1}{m}\sum_{i=1}^{m} \mathrm{e}^{-\frac{\alpha-\beta}{2}\|X_i\|_F^2} \to (\frac{\beta}{\alpha})^{N/4}. \tag{4.5.5}$$

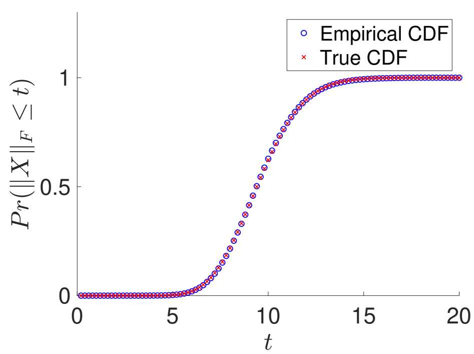


(a) Scheme E (4.3.1) on $(\mathcal{S}_+^{n,p}, g_E)$ with $\Delta t = 0.001$ and $\beta = 0.5$. The error between two CDFs is $KS = 0.0096$

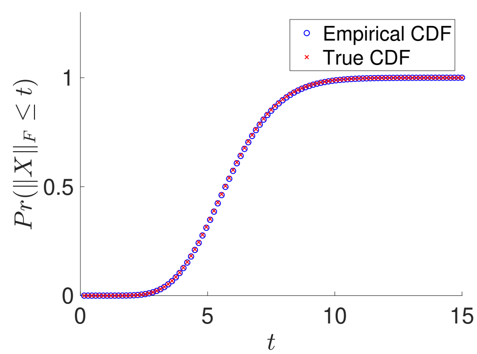


(b) Scheme BW (4.3.5) on $(\mathcal{S}_+^{n,p}, g_{BW})$ with $\Delta t = 0.001$ and $\beta = 0.5$. The error between two CDFs is $KS = 0.0043$.

Figure 4.5.2: Example II: $\mathcal{E}(X) = \mathrm{Tr}(X \log X)$, $n = 5, p = 3$ and manifold dimension is $N = 12$. The empirical CDF is computed by $5E6$ MCMC samples generated after $6E6$ iterations of the Riemannian Langevin Monte Carlo schemes. Both CDFs are evaluated at 100 equally spaced points on $[0, 15]$, and the difference can be measured by the KS statistic (4.5.1).

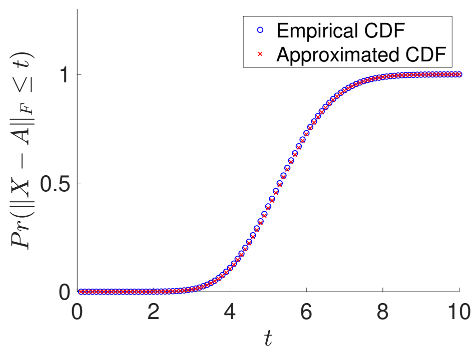


(a) Scheme E (4.3.1) on $(\mathcal{S}_+^{n,p}, g_E)$ with $\Delta t = 0.001$ and $\beta = 0.5$. The error between two CDFs is $KS = 0.006$.

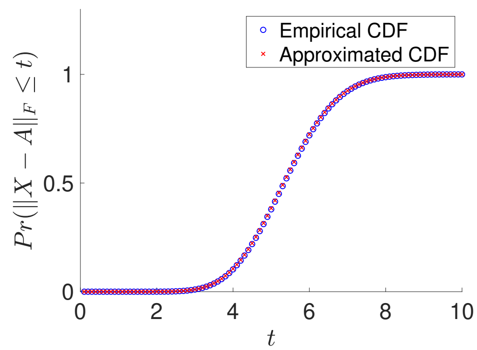


(b) Scheme BW (4.3.5) on $(\mathcal{S}_+^{n,p}, g_{BW})$ with $\Delta t = 0.001$ and $\beta = 0.5$. The error between two CDFs is $KS = 0.0043$.

Figure 4.5.3: Example II: $\mathcal{E}(X) = \mathrm{Tr}(X \log X)$, $n = 10, p = 2$ and manifold dimension is $N = 19$. The empirical CDF is computed by $5E6$ MCMC samples generated after $6E6$ iterations of the Riemannian Langevin Monte Carlo schemes. Both CDFs of scheme E and scheme BW are evaluated at 100 equally spaced points on $[0, 20]$ and $[0, 15]$,respectively, and the difference can be measured by the KS statistic (4.5.1).



(a) Scheme E (4.3.1) on $(\mathcal{S}_+^{n,p}, g_E)$ with $\Delta t = 0.001$ and $\beta = 0.4$. The error between two CDFs is $KS = 0.0084$.



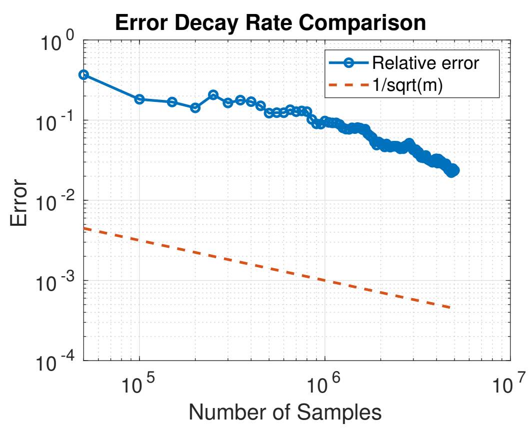
(b) Scheme BW (4.3.5) on $(\mathcal{S}_+^{n,p}, g_{BW})$ with Δt =2E-7 and $\beta = 0.4$. The error between two CDFs is $KS = 0.0052$.

Figure 4.5.4: Example III: $\mathcal{E}(X) = \frac{1}{2}\|X - A\|_F^2$, $n = 5, p = 3$ and manifold dimension is $N = 12$. The nonzero eigenvalues of A are equally spaced between 10000 and 20000. The empirical CDF is computed by $5E6$ MCMC samples generated after $6E6$ iterations of the Riemannian Langevin Monte Carlo schemes. Both CDFs are evaluated at 100 equally spaced points on $[0, 10]$, and the difference can be measured by the KS statistic (4.5.1).

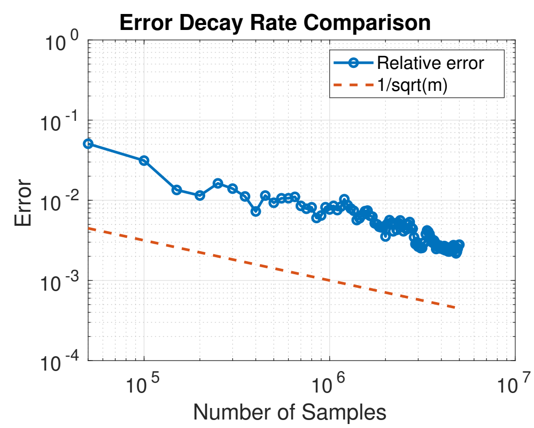
In Figure 4.5.5 we indeed observe the $\mathcal{O}(1/\sqrt{m})$ for the relative error of numerical integration.

4.5.3 A numerical study of the convergence to equilibrium

The general mathematical theory of convergence of a Langevin equation to its equilibrium measure has been well studied; we consider the specific case of the RLE studied here in the companion paper [93]. One particular application of the two Riemannian Langevin Monte Carlo schemes is to use them to numerically study the SDE solutions, e.g., by taking very small time steps, a Riemannian Langevin Monte Carlo scheme approximates the Riemannian Langevin equation on the manifold. We have shown comparison of the Langevin equation on $(\mathcal{S}_+^{n,p}, g_E)$, $(\mathcal{S}_+^{n,p}, g_{BW})$, $\mathbb{R}^{n\times n}$ in Figure 4.5.6, in which we can see interesting differences between two metrics. With all three figures in Figure 4.5.6, we can see that the SDE on $(\mathcal{S}_+^{n,p}, g_{BW})$ has a much faster convergence to its Gibbs measure than the SDE on $(\mathcal{S}_+^{n,p}, g_E)$.

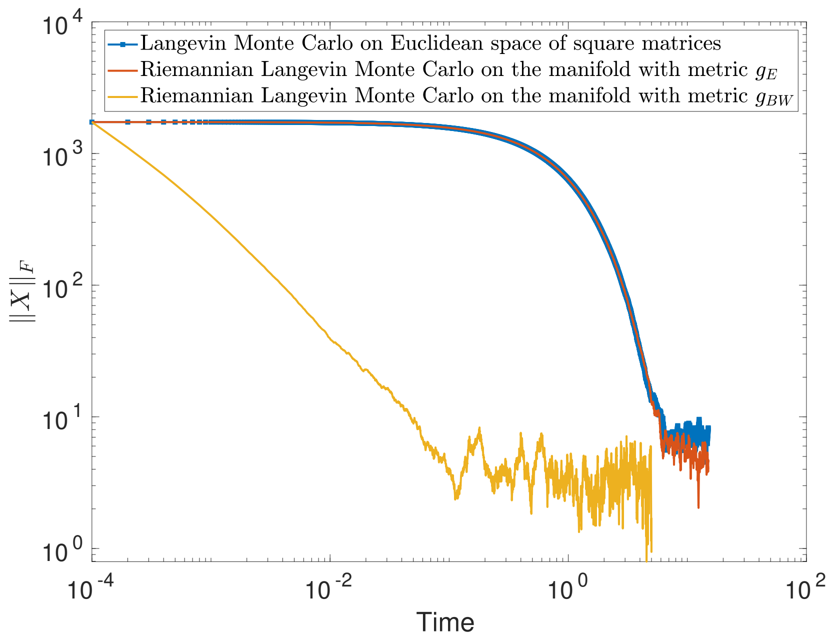


(a) Integration on $(\mathcal{S}_+^{n,p}, g_E)$ via samples generated by Scheme E (4.3.1) with $\Delta t = 0.001$ and $\beta = 0.4$.

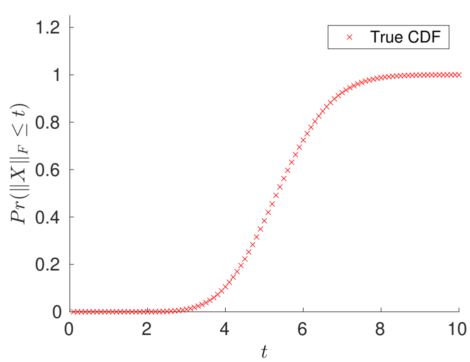


(b) Integration on $(\mathcal{S}_+^{n,p}, g_{BW})$ via samples generated by Scheme BW (4.3.5) with $\Delta t = 0.001$ and $\beta = 0.4$.

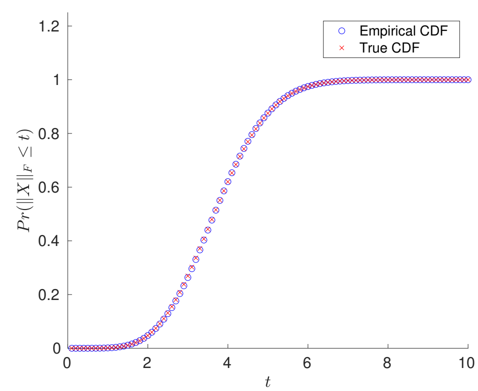
Figure 4.5.5: Convergence rate of the relative error of $\left|\frac{\hat{\mu}_m - \mu}{\mu}\right|$ MCMC integration on the manifold with $n = 10, p = 2$ and dimension $N = 19$. Parameters are $\alpha = 0.75, \beta = 0.4$, for which it is a numerical integration of the function $\mathcal{E}(X) = \frac{1}{2}\|X\|_F^2$ on the manifold $\mathcal{S}_+^{n,p}$. The error shown is the averaged one of 12 independent runs.



(a) Convergence of SDEs to the equilibrium.



(b) The true CDF of $D = \|X\|_F$ on $(\mathcal{S}_+^{n,p}, g_E)$. This implies the blue curve in Figure (a) has not reached equilibrium at Time=1.



(c) The true CDF of $D = \|X\|_F$ on $(\mathcal{S}_+^{n,p}, g_{BW})$, and empirical CDF of 5E6 samples from the Scheme BW after Time=5.

Figure 4.5.6: For $\mathcal{S}_+^{n,p}$ with $n = 5, p = 3$, the dimension is $N = 12$, and $\mathbb{R}^{n\times n}$ has dimension 25. The Gibbs measure is $\mathrm{e}^{-\beta\mathcal{E}(x)}$ with $\beta = 0.4$ and $\mathcal{E} = \frac{1}{2}\|X\|_F^2$. The time step size is $1e-4$. The initial guess is a random PSD matrix of rank-3 with eigenvalues = [1000,1000,1000].

4.6 Conclusion

We have constructed two efficient Riemannian Langevin Monte Carlo schemes for sampling PSD matrices of fixed rank from the Gibbs distribution on the manifold $\mathcal{S}_+^{n,p}$ equipped with two fundamental metrics. We have also provided several examples for which these sampling schemes can be numerically validated.

Chapter 5

RLE in conic programming

5.1 Introduction

5.1.1 Outline

We study Riemannian Langevin equations (RLE) that arise naturally in the study of conic programs. The RLE extends the classical Langevin equation in statistical physics to the setting of a Riemannian manifold. It is of interest in optimization because the RLE provides a mathematical model that unifies the analysis of several algoritheorems. In particular:

1. Canonical Riemannian geometries and Gibbs measures may be associated to both conic programs and certain models of deep learning.

2. RLE have been shown to arise as diffusion limits of certain SGD schemes in machine learning [75].

3. RLE may be used to define Monte Carlo schemes to sample Gibbs measures, extending geometric methods for optimization to Gibbs sampling [94].

The main contributions in this paper are a general result (Theorem 19 below) on

Brownian motion on cones equipped with a canonical metric and an explicit analysis of the Lorentz cone.

5.1.2 The canonical metric

Assume given a regular convex cone $K \subset \mathbb{R}^n$. The *canonical barrier* on K is the convex solution to the Monge-Ampère equation

$$F = \frac{1}{2} \log \det D^2 F, \quad x \in K^o, \tag{5.1.1}$$

with the boundary condition $F(x) \to +\infty$ as $x \to \partial K$. The existence and uniqueness of this solution is a deep result in PDE theory due to Cheng and Yau [21]. We call the Hessian D^2F the *canonical metric*. In what follows, we write $g = D^2F$ for brevity and equip K with the metric g to obtain the Riemannian manifold (K, g).

Hildebrand and Fox have independently established that F is a logaritheoremically homogeneous self-concordant barrier with parameter n [40, 34]. The canonical barrier has several other surprising geometric properties; see [41]. It is identical to the *universal barrier* of Nesterov and Nemirovskii when K is homogeneous. In this case, several examples of F have been computed by Güler [39]. Another barrier, termed the *entropic barrier* has been introduced by Bubeck and Eldan [18]. In general, the relationship between these barriers remains somewhat mysterious. The Monge-Ampère equation (5.1.1) was first studied because of its importance in the Calabi conjecture in geometry and its reappearance in optimization theory is surprising. This paper sheds a different light on these connections using probabilistic methods.

5.1.3 Riemannian Langevin equation

Recall that the Langevin equation associated to the loss function $E : \mathbb{R}^n \to \mathbb{R}$, at inverse temperature $\beta > 0$, is formulated mathematically as the Itô SDE

$$dX_t = -\nabla E(X_t)\,dt + \sqrt{\frac{2}{\beta}}dB_t, \tag{5.1.2}$$

where $\{B_t\}_{t\geq 0}$ denotes standard Brownian motion on $\mathbb{R}^n$.

Given an n-dimensional Riemannian manifold $(\mathcal{M}, g)$ with metric g and a loss function $E : \mathcal{M} \to \mathbb{R}$ we consider the *Riemannian Langevin equation* (RLE)

$$dX_t = -\operatorname{grad} E(X_t)\,dt + dB_t^{g,\beta}. \tag{5.1.3}$$

Both the gradient and the Brownian motion at inverse temperature β are now computed with respect to the Riemannian metric g. While we have stated the RLE as an Itô equation to demonstrate that it is the natural extension of (5.1.2) to $(\mathcal{M}, g)$, equation (5.1.3) is only formal, since SDE on manifolds must be stated in Stratonovich form to ensure coordinate invariance. The more accurate form of (5.1.3) is

[State Stratonovich SDE here and label it eq:rl2]

In particular, the Brownian motion $B_t^{g,\beta}$ on $(\mathcal{M}, g)$ must be defined carefully as discussed below. Let $f \in C^\infty(\mathcal{M})$ be a test function. The infinitesimal generator L of the diffusion (5.1.3) is

$$\mathcal{L}f = -\operatorname{grad} E(f) + \frac{1}{\beta}\Delta f, \quad \text{where} \quad \Delta f = \frac{1}{\sqrt{\det g}}\partial_i(\sqrt{\det g}g^{ij}\partial_j f), \tag{5.1.4}$$

is the Laplace-Beltrami operator, grad denotes the gradient with respect to g, and the volume form is computed in coordinates using $\det g = \det(g_{ij})$.

The Fokker-Planck equation, $\partial_t \rho = L^* \rho$, where the dual is with respect to the volume form of g, takes the form

$$\partial_t \rho = \operatorname{div}\left(\rho \operatorname{grad} F\right), \quad F = E + \frac{1}{\beta}\log \rho. \tag{5.1.5}$$

The free energy, F, is constant in equilibrium and we find the Gibbs density

$$\rho(x) = \frac{1}{Z_\beta} e^{-\beta E(x)}, \quad Z_\beta = \int_{\mathcal{M}} e^{-\beta E(x)} \sqrt{\det g}\, dx. \tag{5.1.6}$$

An important feature of metrics arising in conic programs and deep learning is that the volume $\int_M \sqrt{\det g}\, dx$ may be infinite. Thus, the above calculations are only formal. The existence of Gibbs measures is a subtle problemma.

5.1.4 Riemannian geometries in optimization

The framework of the RLE provides a natural geometric unity between conic programs and deep learning. What changes is the underlying Riemannian manifold $(\mathcal{M}, g)$. Let us illustrate this idea with examples.

First, in the setting of conic programs, the role of Riemannian geometry in interior point methods is implict in Karmarkar's early work on linear programming [54]. A systematic treatment of the underlying Riemannian geometry was provided by Bayer and Lagarias [12]. They showed in particular,

Here we are given a convex polytope $C \in \mathbb{R}^n$

Given a convex barrier and a vector $c \in \mathbb{R}^n$, the conic program $\min_{x \in C} c^T x$ may be solved by taking the $\theta \to \infty$ limit of the *central path*

$$x(\theta) = \operatorname{argmin}_{s \in C}\{F(s) + \theta c^T s\}. \tag{5.1.7}$$

Further, $x(\theta)$ above is the solution to the Riemannian gradient flow

$$\frac{dx}{d\theta} = -\mathrm{grad}_g(c^T x), \quad g = D^2 F, \quad x(0) = \mathrm{argmin} F. \tag{5.1.8}$$

That is, the Hessian of the barrier F provides the underlying Riemannian metric. We now see immediately that the RLE is the natural stochastic dynamical system corresponding to the gradient flow (5.1.8).

Riemannian metrics have also been extensively used in geometric deep learning [16]. A matrix model that allows a comparison between deep learning and conic programs is the deep linear network [5, 6, 27]. The training space for a network of depth N is the product space of $d \times d$ matrices $\mathbb{M}_d^N$. Given $\mathbf{W} = (W_N, W_{N-1}, \ldots, W_1)$ the observable is the product $V = W_N W_{N-1} \cdots W_1$. Learning problemmas like matrix completion may be modeled as a Euclidean gradient descent for a cost function $L(\mathbf{W}) := E(V)$. Then for suitable initial conditions, the Euclidean gradient flow, $\dot{W}_i = -\nabla_{W_i} L(W)$, $1 \leq i \leq N$ corresponds to the Riemannian gradient flow, $\dot{V} = -\mathrm{grad}_g E(V)$, where the metric g acts by

$$g(Z_1, Z_2) = \mathrm{Tr}(A_N^{-1}(Z_1) Z_2), \quad A_N(Z) = \frac{1}{N} \sum_{i=1}^{N} (WW^T)^{\frac{N-i}{N}} Z (W^T W)^{\frac{i}{N}}. \tag{5.1.9}$$

In order to explore the nature of the Riemannian Langevin equation in optimization, we must understand Brownian motion on Riemannian manifolds like those above. This is a problemma of some depth. We illustrate this by computing explicit expressions for Brownian motion in some fundamental cones, using expressions for the barrier from [39].

5.1.5 A comparison of SGD with RLE

Stochastic gradient descent (SGD) schemes in machine learning typically arise as follows. An empirical loss function $E : \mathbb{R}^n \to \mathbb{R}$, for a training parameter x, is defined through a finite sum $E(x) = \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i(x)$, where $\varepsilon_i(x) = \varepsilon(x; y_i, z_i)$ denotes a loss function evaluated on a finite set of training data $\{(y_i, z_i)\}_{i=1}^{N}$. The loss function is minimized using the stochastic gradient descent scheme

$$x^{k+1} = x^k - \gamma_k \nabla \varepsilon_{i_k}(x^k), \tag{5.1.10}$$

where i_k is chosen randomly from the set $\{1, \ldots, N\}$ and γ_k is a time step.

Several variants of SGD have been explored since the classic work of Robbins and Monro [81]. What is different in modern maching learning is the large size of n and N and the protocol for the learning rate γ. Diffusion limits of SGD replace the discrete iteration above with stochastic differential equations (SDE); these SDE depend on the manner in which $\gamma_k \to 0$, $n \to \infty$ and $N \to \infty$. Some examples of this approach are the stochastic modified equation (SME) proposed in [63], the variational analysis using Kullback-Leibler divergence proposed in [67], and *homogenized* SGD (HSGD) defined in [79].

The RLE provides a dynamical system to the study the Gibbs measure associated to F, whereas SGD schemes seek the minimum of E. In algoritheoremic terms, the Euler-Murayama discretization of the RLE provides a Riemannian Langevin Monte Carlo (RLMC) scheme to sample from the Gibbs measure. Despite the obvious difference between optimization and sampling, these techniques are closely related. When $\beta \to \infty$ and E has a unique global minimum at $x_* \in \mathcal{M}$, we expect the Gibbs measures to concentrate at x_* as $\beta \to \infty$ with rigorous asymptotics provided by the Bakry-Emery technique and large deviations theory.

In practice, the design and analysis of RLMC algoritheorems requires an explicit understanding of the underlying Riemannian geometry.

5.2 Brownian motion and conic programs

Manifold-valued Brownian motion may be defined in several ways [43]. We use the following definition in this note: an $\mathcal{M}$-valued semimartingale $\boldsymbol{X}_t$ is called a Brownian motion on $\mathcal{M}$, with temperature $T = \frac{1}{\beta}$, if for any $f \in C^\infty(\mathcal{M})$,

$$f(\boldsymbol{X}_t) = f(\boldsymbol{X}_0) + \frac{1}{\beta}\int_0^t \Delta f(\boldsymbol{X}_s)\mathrm{d}s + \text{a local martingale}. \tag{5.2.1}$$

We denote Brownian motion on $(\mathcal{M}, g)$ at temperature $T = \frac{1}{\beta}$ by $\boldsymbol{B}_t^{g,\beta}$.

Proposition 5. *The quadratic variation process of $f(\boldsymbol{B}_t^{g,\beta})$ for $f \in C^\infty(\mathcal{M})$ is*

$$[f(\boldsymbol{B}^{g,\beta})]_t = \frac{1}{\beta}\int_0^t |\mathrm{grad} f(\boldsymbol{B}_s^{g,\beta})|_g^2 \mathrm{d}s \tag{5.2.2}$$

where $|\mathrm{grad} f|^2 := g^{ij}\partial_i f \partial_j f$.

Corollary 7. *The covariation process of $f_1(\boldsymbol{B}_t^{g,\beta})$ and $f_2(\boldsymbol{B}_t^{g,\beta})$ for $f_1, f_2 \in C^\infty(\mathcal{M})$ is*

$$[f_1(\boldsymbol{B}^{g,\beta}), f_2(\boldsymbol{B}^{g,\beta})]_t = \frac{1}{\beta}\int_0^t (\mathrm{grad} f_1 \cdot \mathrm{grad} f_2)(\boldsymbol{B}_s^{g,\beta})\mathrm{d}s \tag{5.2.3}$$

where $\mathrm{grad} f_1 \cdot \mathrm{grad} f_2 := g^{ij}\partial_i f_1 \partial_j f_2$.

Proposition 5 allows us to analyze Brownian motion through a careful choice of coordinate functions. We will choose f such that $\Delta f = 0$ and $|\mathrm{grad} f|^2 = 1$, so that $f(\boldsymbol{B}_t^{g,\beta}) - f(\boldsymbol{B}_0^{g,\beta})$ has the same law as an $\mathbb{R}$-valued standard Brownian motion.

Let us now assume $\mathcal{M}$ is a regular convex cone $K \subset \mathbb{R}^n$, let F denote its canonical barrier, and equip K with the Hessian metric

$$g_{ij} = \frac{\partial^2 F}{\partial x^i \partial x^j}. \tag{5.2.4}$$

Theorem 19. *Consider Brownian motion $\boldsymbol{B}_t^{g,\beta}$ on (K, g). The process $\frac{\sqrt{\beta}}{\sqrt{n}}(F(\boldsymbol{B}_t^{g,\beta}) - F(\boldsymbol{B}_0^{g,\beta}))$ has the same law as a standard Brownian motion on the line.*

Proof. We will use the logaritheoremic homogeneity of F and the Monge-Ampère equation to establish the identities

$$\text{grad}F = -x, \quad |\text{grad}F|^2 = n \quad \text{and} \quad \Delta F = 0. \tag{5.2.5}$$

Theorem 19 follows immediately from these identities.

The first identity uses logaritheoremic homogeneity. Since $F(\lambda x) = F(x) - n \log \lambda$ for $\lambda \in \mathbb{R}^+$, $x \in K$, we may differentiate with respect to λ and set $\lambda = 1$ to find

$$x^i \partial_i F(x) = -n. \tag{5.2.6}$$

Next, the differential of F with respect to x^j is $\partial_j F(x) = -x^i \partial_i \partial_j F(x) = -g_{ij} x^i$ by equation (5.2.4). Thus, each component of the gradient of F is

$$(\text{grad}\, F)^k = g^{jk} \partial_j F = -x^k. \tag{5.2.7}$$

This proves the first identity in equation (5.2.5). It immediately follows that

$$|\text{grad}F|^2 = (g^{ij} \partial_i F) \partial_j F = -x^j \partial_j F = n. \tag{5.2.8}$$

Finally, we show that $\Delta F = 0$ as follows

$$\begin{aligned}\Delta F = \text{div}(\text{grad}F) &= -\frac{1}{\sqrt{\det g}} \partial_i (\sqrt{\det g}\, x^i) \\ &= -\partial_i (x^i) - x^i \partial_i (\frac{1}{2} \log \det g) = -n - x^i \partial_i F = 0,\end{aligned}$$

where we have used the Monge-Ampère equation (5.1.1) and equation (5.2.8). $\square$

The above theorem sheds new light on the mysterious reappearance of the Cheng-Yau metric in optimization theory. Let us understand it better with examples.

5.3 Brownian motion examples

5.3.1 Positive orthant

Denoted by $\mathbb{R}^n_+ = \{x \in \mathbb{R}^n | x^i > 0, i = 1, ..., n\}$ the positive orthant. The canonical barrier and its Hessian metric are

$$F(x) = -\sum_{i=1}^{n} \log x^i, \quad g = \sum_{i=1}^{n} \frac{\mathrm{d}x^i \mathrm{d}x^i}{(x^i)^2}. \tag{5.3.1}$$

Then for each choice of coordinate, we have the identity in law

$$\log x^i(\boldsymbol{B}_t^{g,\beta}) - \log x^i(\boldsymbol{B}_0^{g,\beta}) = \frac{1}{\sqrt{\beta}} B_t^i \qquad i = 1, ..., n \tag{5.3.2}$$

where $\{B_t^i\}_{i=1}^n$ are n independent standard Brownian motion on $\mathbb{R}$.

5.3.2 Cube

Next we consider a convex *set*, the cube $B_n = (0,1)^n$. We find that

$$F(x) = -\sum_{i=1}^{n} \log \frac{\sin(\pi x^i)}{\pi}, \quad g = \sum_{i=1}^{n} \pi^2 \frac{\mathrm{d}x^i \mathrm{d}x^i}{\sin^2(\pi x^i)}. \tag{5.3.3}$$

Similar calculations yield the identity in law

$$\log(\tan(\frac{\pi x^i(\boldsymbol{B}_t^{g,\beta})}{2})) - \log(\tan(\frac{\pi x^i(\boldsymbol{B}_0^{g,\beta})}{2})) = \frac{1}{\sqrt{\beta}} B_t^i \tag{5.3.4}$$

where $\{B_t^i\}_{i=1}^n$ are n independent standard Brownian motions on $\mathbb{R}$.

5.3.3 Lorentz cone

A deeper example is provided by the Lorentz cone

$$K_{n+1} = \left\{ x \in \mathbb{R}^{n+1} | (x^0) \geq \sqrt{\sum_{i=1}^{n} (x^i)^2} \right\}, \tag{5.3.5}$$

where $x = (x^0, \ldots, x^n)$. The canonical barrier F on K_{n+1} is given by

$$F(x) = -\frac{n+1}{2} \log(x^T Ax) + \frac{n+1}{2} \log(n+1), \quad A = \mathrm{diag}(1, -1, -1, \cdots, -1).$$

The metric g, its inverse, and volume form are as follows

$$g_{ij} = -\frac{n+1}{x^T Ax} \big(A_{ij} - 2\frac{A_{ik} x^k A_{jl} x^l}{x^T Ax} \big), \tag{5.3.6}$$

$$g^{ij} = -\frac{1}{n+1} \left((x^T Ax) B^{ij} - 2x^i x^j \right), \tag{5.3.7}$$

$$\sqrt{\det(g_{ij})} = \left(\frac{n+1}{x^T Ax} \right)^{\frac{n+1}{2}} = \exp(F), \tag{5.3.8}$$

where $B = A^{-1}$ is the inverse of A. In our case we have $B = A$, but these matrices are conceptually distinct.

We characterize Brownian motion on K_{n+1} using the auxiliary functions

$$f_b(x) = \frac{\sqrt{n+1}}{2} \log \frac{(b^T x)^2}{x^T Ax}, \quad b \in L^+_{n+1}. \tag{5.3.9}$$

Here we have introduced the light-cone

$$L^+_{n+1} = \left\{ b \in \mathbb{R}^{n+1} | b^0 > 0, b^T Bb = 0 \right\}.$$

Pick n vectors $\{b_i\}_{i=1}^n \subset L_{n+1}^+$ and define the $n+1$ functions

$$f^0 = \frac{1}{\sqrt{n+1}} F, \quad \text{and} \quad f^i = f_{b_i} \quad i = 1, \ldots, n.$$

We choose a drift and covariance tensor as follows

$$\mu^0 = 0, \quad \mu^i = \frac{n-1}{2\sqrt{n+1}}, \quad i = 1, \ldots, n. \tag{5.3.10}$$

$$\Sigma^{ij} = 1 - b_i^T B b_j \exp\left(-\frac{f_t^i + f_t^j}{\sqrt{n+1}}\right), \quad i, j = 1, \ldots n. \tag{5.3.11}$$

Finally, set $\Sigma^{00} = 1$ and $\Sigma^{ij} = 0$ when exactly one of the indices is zero.

Theorem 20. *Denote by $\boldsymbol{B}_t$ Brownian motion on (K_{n+1}, g) with $\beta = 2$. The stochastic processes $f_t^i := f^i(\boldsymbol{B}_t)$ satisfy the Itô SDE*

$$\mathrm{d}f_t^i = \mu^i \mathrm{d}t + \sigma_j^i \mathrm{d}B_t^j, \quad f_0^i = f^i(\boldsymbol{B}_0) \qquad i = 0, 1, ..., n. \tag{5.3.12}$$

In particular, each f_t^i is itself identical in law with a Brownian motion with constant drift.

Proof. We only need to check that

$$\frac{1}{2}\Delta f^i = \mu^i, \quad \text{and} \quad \mathrm{grad} f^i \cdot \mathrm{grad} f^j = \Sigma^{ij}, \quad i, j = 0, \ldots, n. \tag{5.3.13}$$

First, when $i = j = 0$, this is just the claim of Theorem 19. When exactly one of the indices is zero, we use equation (5.2.5) and the fact that $\frac{(b^T x)^2}{x^T A x}$ is a homogeneous polynomial of order 0 to obtain

$$\mathrm{grad} f^0 \cdot \mathrm{grad} f^i = \frac{1}{\sqrt{n+1}} \mathrm{grad}\, F(\mathrm{d}f^i) = -\frac{1}{\sqrt{n+1}} x^k \partial_k f^i = 0. \tag{5.3.14}$$

Finally, consider the case when both i and j are space-like. We start with the following property: for $b, b' \in L^+_{n+1}$,

$$\begin{aligned}\operatorname{grad}\log(b^Tx)\cdot\operatorname{grad}\log(b'^Tx) &= -\frac{1}{n+1}\left((x^TAx)B^{ij}-2x^ix^j\right)\frac{b'_i}{b^Tx}\frac{b_j}{b^Tx}\\ &= \frac{1}{n+1}\left(2-(b^TBb')\frac{(x^TAx)}{(b^Tx)(b'^Tx)}\right).\end{aligned} \tag{5.3.15}$$

Using the fact that $\log(b^Tx) = \frac{1}{\sqrt{n+1}}(f_b - f^0)$ and $\operatorname{grad}f_b\cdot\operatorname{grad}f^0 = 0$, we have

$$\begin{aligned}\operatorname{grad}f_b\cdot\operatorname{grad}f_{b'} &= (n+1)\operatorname{grad}\log(b^Tx)\cdot\operatorname{grad}\log(b'^Tx) - |\operatorname{grad}f^0|^2\\ &= 1-(b^TBb')\frac{(x^TAx)}{(b^Tx)(b'^Tx)}\\ &= 1-(b^TBb')\exp\left(-\frac{f_b+f_{b'}}{\sqrt{n+1}}\right).\end{aligned} \tag{5.3.16}$$

In particular, we find that $|\operatorname{grad}f_b|^2 = 1$ because $b^TBb = 0$ when $b \in xL^+_{n+1}$.

The proof of the first identity in equation (5.3.13) is a computation:

$$\begin{aligned}\left(\text{grad }\log(b^Tx)\right)^i &= -\frac{1}{n+1}((x^TAx)B^{ij}-2x^ix^j)\frac{b_i}{b^Tx} = \frac{1}{n+1}(2x^i - \frac{x^TAx}{b^Tx}B^{ij}b_j)\\ \Rightarrow \Delta\log(b^Tx) &= \frac{1}{\sqrt{\det g}}\partial_i\left(\sqrt{\det g}(\text{grad }\log(b^Tx))^i\right)\\ &= \partial_i\left((\text{grad }\log(b^Tx))^i\right) + \left(\text{grad }\log(b^Tx)\right)^i\partial_iF\\ &= \partial_i\left(\frac{1}{n+1}(2x^i - \frac{x^TAx}{b^Tx}B^{ij}b_j)\right) + (\text{grad }F)^i\partial_i\log(b^Tx)\\ &= \frac{1}{n+1}(2(n+1)-2)-1 = \frac{n-1}{n+1}.\end{aligned} \tag{5.3.17}$$

Thus, finally we have

$$\Delta f_b = \Delta(\sqrt{n+1}\log(b^Tx)+f^0) = \frac{n-1}{\sqrt{n+1}}.$$

$\square$

5.4 Additional properties

5.4.1 Divergence of volume

Theorem 21. *Let $K \subset \mathbb{R}^{n+1}$ be a regular convex cone, F be its canonical barrier, and g be the metric generated by F defined on K^o. Then the volume of K^o, defined by the integral* $\mathrm{Vol}(K) := \int_{K^o} \sqrt{\det g}\mathrm{d}x$ *is always infinite.*

Proof. Without loss of generality, assume K^o is in the region $x_0 > 0$, so $K_s = K \cap \{x_0 = s\}$ is non-empty and finite for $s > 0$, and $K = \cup_{s>0} K_s$.

As the canonical barrier F of K^o satisfies Monge-Ampère equation, we have

$$\begin{aligned} \sqrt{\det g} &= \sqrt{\det D^2 F} = \mathrm{e}^F \\ \Leftarrow \mathrm{Vol}(K) &= \int_{K^o} \mathrm{e}^F \mathrm{d}x \,. \end{aligned} \tag{5.4.1}$$

By homogeneity of F:

$$F(x_0, x_1, ..., x_n) = F(1, \frac{x_1}{x_0}, ..., \frac{x_n}{x_0}) - (n+1)\log x_0 \tag{5.4.2}$$

Let $s = x_0$, $y_i = \frac{x_i}{x_0}$ for $i = 1, ..., n$, and $f(y) = F((1, y))$, the volume of K has the following expression:

$$\begin{aligned} \mathrm{Vol}(K) &= \int_{K^o} \mathrm{e}^F \mathrm{d}x \\ &= \int_{\mathbb{R}_+ \times K_1} \mathrm{e}^{f(y)-(n+1)\log s} s^n \mathrm{d}s\mathrm{d}y \\ &= \Big(\int_0^\infty \frac{1}{s}\mathrm{d}s \Big) \times \Big(\int_{K_1} \mathrm{e}^{f(y)}\mathrm{d}y \Big) \,. \end{aligned} \tag{5.4.3}$$

□

The first factor is clearly divergent, which makes $\mathrm{Vol}(K)$ infinite.

5.4.2 Mean curvature of level surface

Theorem 22. *Let $K \subset \mathbb{R}^{n+1}$ be a regular convex cone, F be its canonical barrier, and g be the metric generated by F defined on K^o. For $\alpha \in \mathbb{R}$, let $\mathcal{G}_\alpha = \{x \in K | F(x) = \alpha\}$ be the level surface of F, then $\mathcal{G}_\alpha$ has mean curvature 0.*

Proof. As $\mathcal{G}_\alpha$ has codimension 1, the mean curvature $\boldsymbol{H}$ should be proportional to $\operatorname{grad} F = -\boldsymbol{x}$, i.e. $\boldsymbol{H} = c(x)\boldsymbol{x}$, and we can evaluate $c(x)$ by evaluating $H(F)$, so $c(x) = \frac{\boldsymbol{H}(F)}{\boldsymbol{x}(F)}$.

Let $\{\boldsymbol{u}_i\}_{i=1}^n$ be an orthonormal basis of $T\mathcal{G}_\alpha$, by definition we have

$$
\begin{aligned}
\boldsymbol{H}(F) =& \sum_{i=1}^n (\nabla_{\boldsymbol{u}_i} \boldsymbol{u}_i)(F) \\
=& \sum_{i=1}^n \left[\boldsymbol{u}_i(\boldsymbol{u}_i(F)) - \nabla^2 F(\boldsymbol{u}, \boldsymbol{u})\right] \\
=& \sum_{i=1}^n \boldsymbol{u}_i(\boldsymbol{u}_i(F)) - \Delta F + \nabla^2 F(\frac{1}{\sqrt{n+1}}\boldsymbol{x}, \frac{1}{\sqrt{n+1}}\boldsymbol{x}),
\end{aligned}
$$

where we used the fact that $\frac{1}{\sqrt{n+1}}\boldsymbol{x}$ along with $\{\boldsymbol{u}_i\}_{i=1}^n$ forms an orthonormal basis of K^o. As $\boldsymbol{u}_i(F) = 0$, $\Delta F = 0$, we have

$$
\begin{aligned}
\boldsymbol{H}(F) =& \frac{1}{n+1} \nabla^2 F(\boldsymbol{x}, \boldsymbol{x}) \\
=& \frac{1}{n+1}\left[\boldsymbol{x}(\boldsymbol{x}(F)) - (\nabla_{\boldsymbol{x}} \boldsymbol{x})(F)\right] \\
=& \frac{1}{n+1}\left[\boldsymbol{x}(\boldsymbol{x}(F)) - \langle \nabla_{\boldsymbol{x}} \boldsymbol{x}, \operatorname{grad} F\rangle\right] \\
=& \frac{1}{n+1}\left[\boldsymbol{x}(\boldsymbol{x}(F)) + \langle \nabla_{\boldsymbol{x}} \boldsymbol{x}, \boldsymbol{x}\rangle\right].
\end{aligned} \tag{5.4.4}
$$

Since $\boldsymbol{x}(F) = -n - 1$ is a constant, $\boldsymbol{x}(\boldsymbol{x}(F)) = 0$. Also, as $\langle \boldsymbol{x}, \boldsymbol{x}\rangle = n + 1$ is a constant, $\langle \nabla_{\boldsymbol{x}} \boldsymbol{x}, \boldsymbol{x} \rangle = \frac{1}{2}\boldsymbol{x}(\langle \boldsymbol{x}, \boldsymbol{x} \rangle) = 0$, so we conclude that $\boldsymbol{H}(F) = 0$, therefore $\boldsymbol{H} = 0$. $\square$

Chapter 6

RLE in Deep Linear Network

6.1 Introduction

6.1.1 Deep Linear Network

Deep neutral network has become the central technique of artificial intelligence, while mathematical theories on deep neutral network and explanations of its effectiveness are not satisfactory at this stage. To approach rigorous analysis of deep neutral network, possible simplified models include two-layer neutral network [64] and deep linear network (DLN) [4]. DLN is the deep neutral network with all activation functions being identity map, which is a highly simplified model but still preserves properties of deep neutral networks, e.g. number of parameters can be extremely larger than data size, so the study of DLN can help understanding overparametrization problem in deep neutral networks.

Overparametrization or overfitting problems arises in classical regression models, and methods including ridge regression and Lasso regression are known as explicit regularization to mitigate this problem. In practice, however, deep neutral network seems to avoid overfitting problem without explicit regularization, which is called implicit regularization. In the work [28] the implicit regularization in DLN is interpreted as the effect of state space volume. The volume of state space is based on the hidden Riemannian structure in the gradient flow of DLN [7].

In this paper, we obtain the same Riemannian structure of DLN in a pure geometric way. We also provide description of Riemannian Langevin dynamics on the balanced manifold of DLN, explaning the effect of state space volume on the choice of minimizer of objective function. The effect of state space volume on optimization problems is common when stochastic gradient descent algorithms are applied. From the dynamical point of view, projection of Brownian motion from higher dimensional parameter space to lower dimension space produces a repulsion given by mean curvature of fiber. In view of stationary distribution, the volume of state space provides a further factor on the Gibbs distribution, resulting in implicit regularization.

6.1.2 Balanced manifold of DLN

An N-layer DLN with width d is described by N of $d \times d$ real valued matrices $\{W_i\}_{i=1}^{N}$. Given input $z_0 \in \mathbb{R}^d$, the output of each layer is defined by $z_i = W_i z_{i-1}$ for $i = 1, ..., N$ and the overall output $z_N = W z_0$ where $W = W_N W_{N-1} \cdots W_2 W_1$. Optimization problem of DLN can be formalized as

$$\min_{\underline{W}=(W_N,...,W_1)} E(W) \tag{6.1.1}$$

for some energy function $E \in C^2(\mathbb{R}^{d\times d}$.

A common choice of the energy function $E(W)$ is

$$E(W) = \sum_j ||y^{(j)} - W z^{(j)}||^2 = \sum_j ||y^{(j)} - W_N W_{N-1} \cdots W_2 W_1 z^{(j)}||^2 \tag{6.1.2}$$

for dataset $\{(z^{(j)}, y^{(j)})\}_j$.

When applying gradient flow on $E(W)$, it is observed [4] that $G_i = W_{i+1}^T W_{i+1} - W_i W_i^T$ for $i = 1, ..., N-1$ are conserved quantities, and the variety

$$\left\{\underline{W} \in (\mathbb{R}^{d\times d})^N \middle| W_{i+1}^T W_{i+1} = W_i W_i^T, i = 1, ..., N-1\right\} \tag{6.1.3}$$

foliated by rank r of W_i's gives balanced manifolds. In this paper, we restrict our attention to the balanced manifold with full rank and distinct eigenvalues. Define:

$$\begin{aligned}\mathcal{M} =& \Big\{\underline{W} \in (\mathbb{R}^{d\times d})^N \Big| W_{i+1}^T W_{i+1} = W_i W_i^T, , i = 1, ..., N-1, \\ & W_i \text{ is full rank with distinct singular values}, i = 1, ..., N\Big\}.\end{aligned} \tag{6.1.4}$$

With each W_i being full rank, the product W is also full rank. Let $\mathcal{N} = \{W \in \mathbb{R}^{d\times d} | W \text{ is full rank with distinct singular values}\}$ be the image of $\mathcal{M}$ under $\phi : (W_1, ..., W_N) \to W = W_N W_{N-1} \cdots W_1$, we will explore the geometric structures of $\mathcal{M}$ and $\mathcal{N}$, regarding $\mathcal{M}$ as a submanifold of $(\mathbb{R}^{d\times d})^N$ and $\mathcal{N}$ a quotient manifold of $\mathcal{M}$.

6.2 Main results

The main result of this paper are two theorems describing the effect of mean curvature of fiber $\phi^{-1}(W)$ on the evolution of W. The first theorem provides the volume of $\phi^{-1}(W)$, and the second theorem describes the effect of mean curvature in the optimization of $E(W)$.

Theorem 29. For each $W \in \mathcal{N}$ with SVD $W = Q_N \Sigma Q_0^T,\ Q_N, Q_0 \in \mathrm{O}(d), \Sigma = \mathrm{diag}(\sigma_1, ..., \sigma_d)$, the volume of $\phi^{-1}(W)$ is given by

$$\mathrm{vol}(\phi^{-1}(W)) = c_d^{N-1}\sqrt{\frac{\mathrm{van}(\Sigma^2)}{\mathrm{van}(\Sigma^{\frac{2}{N}})}} = c_d^{N-1}\prod_{1\leq i<j\leq d}\sqrt{\frac{\sigma_i^2-\sigma_j^2}{\sigma_i^{\frac{2}{N}}-\sigma_j^{\frac{2}{N}}}} \tag{6.2.1}$$

where $c_d = \mathrm{vol}(\mathrm{O}(d)) = 2^{\frac{1}{2}d(d+3)}\prod_{r=1}^d \frac{\pi^{\frac{r}{2}}}{\Gamma(\frac{r}{2})}$ is the volume of orthogonal group $\mathrm{O}(d)$ [80].

Remark 30. This theorem should be compared with Theorem 1.1 in [28] where the volume form of $\mathcal{N}$ is computed. The product of $\mathrm{vol}(\phi^{-1}(W))$ and the volume form in [28] gives a measure on $\mathcal{N}$ that for each $U \subset \mathcal{N}$ evaluates the volume of $\phi^{-1}(U) \subset \mathcal{M}$.

Once the volume is given, the mean curvature of fiber $\phi^{-1}(W)$ is given by $\boldsymbol{H}(W) = -\mathrm{grad}\log\mathrm{vol}(\phi^{-1}(W))$ [77]. Therefore, if $\underline{W}_t$ solves the Riemannian Langevin equation with energy function $\tilde{E}(\underline{W}) = E(\phi(\underline{W}))$, $W_t = \phi(\underline{W}_t)$ solves Riemannian Langevin equation with Helmholtz free energy function $F(W) = E(W) - \frac{1}{\beta}\log\mathrm{vol}(\phi^{-1}(W))$.

Theorem 31. If $\underline{W}_t$ solves RLE

$$\mathrm{d}\underline{W}_t = -\mathcal{D}(E\circ\phi)(\underline{W}_t)\mathrm{d}t + \mathrm{d}B_t^{g,\beta} \tag{6.2.2}$$

where $B_t^{g,\beta}$ is the Brownian motion on $\mathcal{M}$, with metric g induced by the metric $\sum_{i=1}^N \mathrm{Tr}(\mathrm{d}W_i \mathrm{d}W_i^T)$ in Euclidean space. $\mathcal{D}$ is the differential operator, and $\mathcal{D}(E\circ\phi)$ is the differential of $E\circ\phi$ with respect to Nd^2 variables.

Then $W_t = \phi(\underline{W}_t)$ solves the following RLE

$$\mathrm{d}W_t = -\mathrm{grad}_{g_{\mathcal{N}}} F(W_t)\mathrm{d}t + \mathrm{d}B_t^{g_{\mathcal{N}},\beta} \tag{6.2.3}$$

where $F(W) = E(W) - \frac{1}{\beta}\log\mathrm{vol}(\phi^{-1}(W))$, and $g_{\mathcal{N}}$ is defined as

$$\begin{aligned} g_{\mathcal{N}} =& \mathrm{Tr}(\mathcal{A}^{-1}(\mathrm{d}W)\mathrm{d}W^T) \\ \mathcal{A}(Z) =& \sum_{i=1}^{N}(WW^T)^{\frac{N-i}{N}} Z (W^TW)^{\frac{i-1}{N}} \end{aligned} \tag{6.2.4}$$

In explicit form, for $W_t = Q_N \sum_{k=1}^d \lambda_k^N e_k e_k^T Q_0^T$, W_t solves the following SDE:

$$\begin{aligned} \mathrm{d}W_t =& -\sum_{i=1}^{N}(W_tW_t^T)^{\frac{N-i}{N}} \mathcal{D}E(W_t)(W_t^TW_t)^{\frac{i-1}{N}}\mathrm{d}t \\ &+ \frac{1}{\beta}\sum_{k=1}^{d}\sum_{l\neq k} Q_N\left(\frac{N\lambda_k^{2N-1}}{\lambda_k^{2N}-\lambda_l^{2N}} - \frac{\lambda_k}{\lambda_k^2-\lambda_l^2}\right)\lambda_k^{N-1} e_k e_k^T Q_0^T \mathrm{d}t \\ &+ \sqrt{\frac{2}{\beta}} Q_N \begin{pmatrix} \sqrt{N}\lambda_1^{N-1}\mathrm{d}B_t^{1,1} & \sqrt{\frac{\lambda_1^{2N}-\lambda_2^{2N}}{\lambda_1^2-\lambda_2^2}}\mathrm{d}B_t^{1,2} & \cdots \\ \sqrt{\frac{\lambda_1^{2N}-\lambda_2^{2N}}{\lambda_1^2-\lambda_2^2}}\mathrm{d}B_t^{1,2} & \sqrt{N}\lambda_1^{N-1}\mathrm{d}B_t^{1,1} & \ddots \\ \vdots & \ddots & \ddots \end{pmatrix} Q_0^T \\ &+ \frac{1}{\beta}\sum_{k=1}^{d} Q_N(N-1)\lambda_k^{N-2}Q_0^T\mathrm{d}t \end{aligned} \tag{6.2.5}$$

where $\mathcal{D}E$ here is the differential of E with respect to d^2 variables, $\{B_t^{k,l}\}_{1\leq k,l\leq d}$ are d^2 independent $\mathbb{R}$-valued standard Brownian motions.

In the explicit form of RLE above, the first two terms are corresponding to the gradient of energy and entropy term in $\mathcal{F}$, and the last two terms are the Brownian motion on the manifold $(\mathcal{N}, g_{\mathcal{N}})$.

6.3 Geometric analysis of $\mathcal{M}$

We first provide a theorem about the structure of $\mathcal{M}$. The constraint $W_{i+1}^T W_{i+1} = W_i W_i^T$ gives strong restriction on the eigenvalues and eigenvectors of matrices $W_N, ..., W_1$. We have the following theorem providing a parametrization of $\mathcal{M}$

Theorem 32. For each $\underline{W} \in \mathcal{M}$, there exists a diagonal matrix $\Lambda \in \mathbb{R}_+^d$ and $N+1$ matrices $\{Q_i\}_{i=0}^N \subset \mathrm{O}(d)$, such that

$$W_i = Q_i \Lambda Q_{i-1}^T \tag{6.3.1}$$

Proof. Let the SVD of each W_i be $W_i = Q_i \Lambda_i P_i^T$. The condition $G_i = W_{i+1}^T W_{i+1} - W_i W_i^T$ requires that

$$Q_i \Lambda_i^2 Q_i^T = P_{i+1} \Lambda_{i+1}^2 P_{i+1}^T \qquad i = 1, ..., N-1\,, \tag{6.3.2}$$

so a possible solution is that $\Lambda_i = \Lambda_{i+1}$ and $P_{i+1} = Q_i$ for all $i = 1, ..., N-1$. □

Remark 33. In the region where singular values of W are all distinct, the map $\mathbb{R}_+^d \otimes (\mathrm{O}(d))^{N+1} \to \mathcal{M}$ is locally bijective. Under operation $Q_i \to Q_i E$ for E being a permutation matrix with entries being ± 1 and $i = 1, ..., N-1$, the image $W \in \mathcal{M}$ is the same, and the group generated by these operations is finite. Up to these finite operations, we regard the SVD: $W_i = Q_i \Lambda Q_{i-1}^T$ as uniquely determined.

In the case W has repeated eigenvalues, there are infinitely many choices of SVD.

From theorem 32, the constraints for $\underline{W} \in \mathcal{M}$ is transformed into the restriction on the SVD decomposition of each W_i. As all W_i share the same spectrum Λ and their $\mathrm{O}(d)$ components cancel each other, it is easy to see that $W = \phi(\underline{W}) = W_N W_{N-1} \cdots W_2 W_1 = Q_N \Lambda^N Q_0^T$, and for given W with distinct singular values, the fiber $\phi^{-1}(W)$ on $\mathcal{M}$ is homeomorphic to $(\mathrm{O}(d))^{N-1}$.

With the parametrization, we are able to analyze the tangent space of $\mathcal{M}$ by differentiating parameters. The following theorem describes the tangent space of $\mathcal{M}$ and provides an orthonormal basis of $\mathfrak{X}(\mathcal{M})$.

Theorem 34. The following vector fields provides an orthonormal basis of $\mathfrak{X}(\mathcal{M})$. Let each W_i be expressed as $W_i = Q_i \Lambda Q_{i-1}^T$, where $\Lambda = \operatorname{diag}(\lambda_1, ..., \lambda_d)$, then for $i = 1, ..., N$, define the following vector fields:

$$L_j(W_i) = \frac{1}{\sqrt{N}} Q_i e_j e_j^T Q_{i-1}^T \,, \tag{6.3.3}$$

for $j = 1, ..., d$ and

$$\begin{aligned} U_p^{k,l}(W_i) =& \sqrt{\frac{1}{N(\lambda_k^2 + \lambda_l^2 - 2\lambda_k \lambda_l \cos \frac{p\pi}{N})}} Q_i \big[(\lambda_k \sin \frac{(i-1)p\pi}{N} - \lambda_l \sin \frac{ip\pi}{N}) e_k e_l^T \\ & + (\lambda_l \sin \frac{(i-1)p\pi}{N} - \lambda_k \sin \frac{ip\pi}{N}) e_l e_k^T \big] Q_{i-1}^T \,, & (6.3.4) \\ U_0^{k,l}(W_i) =& \sqrt{\frac{\lambda_k^2 - \lambda_l^2}{\lambda_k^{2N} - \lambda_l^{2N}}} \lambda_k^{i-1} \lambda_l^{N-i} Q_i e_l e_k^T Q_{i-1}^T \,, & (6.3.5) \\ U_N^{k,l}(W_i) =& \sqrt{\frac{\lambda_k^2 - \lambda_l^2}{\lambda_k^{2N} - \lambda_l^{2N}}} \lambda_k^{N-i} \lambda_l^{i-1} Q_i e_k e_l^T Q_{i-1}^T \,, & (6.3.6) \end{aligned}$$

for $p = 1, ..., N-1$ and $1 \le k < l \le d$.

Remark 35. With the parametrization of $\mathcal{M}$, we can find tangent vector fields of $\mathbb{R}_+^d$ and $\mathbb{O}(d)$ respectively and consider their push forward on $\mathfrak{X}(\mathcal{M})$ to establish basis of $\mathfrak{X}(\mathcal{M})$. Then orthonormalize the basis we will get an orthonormal basis. In the following proof, vector fields $\{l_i\}_{i=1}^d$ (defined below) are the push forward of tangent vector fields on $\mathbb{R}_+^d$ and $\{V_p^{k,l}\}_{1\le k<l\le d, 0\le p\le N}$ are those on $(\mathrm{O}(d))^{N+1}$.

Proof. First consider the case where singular values $\{\lambda_i\}_{i=1}^d$ are all distinct. Using the parametrization of $\mathcal{M}$, we are able to get a basis of $\mathfrak{X}(\mathcal{M})$. For $j = 1, ..., d$, define vector field l_j as

$$l_j(W_p) = Q_p e_j e_j^T Q_{p-1} \qquad\qquad p = 1, ..., N \tag{6.3.7}$$

and for $1 \leq k < l \leq d$, $p = 0, ..., N$, define vector field $V_p^{k,l}$ as

$$\begin{aligned}
V_p^{k,l}(W_p) =& Q_p \frac{-e_k e_l^T + e_l e_k^T}{\sqrt{2}} \Lambda Q_{p-1}^T \\
V_p^{k,l}(W_{p+1}) =& - Q_{p+1} \Lambda \frac{-e_k e_l^T + e_l e_k^T}{\sqrt{2}} Q_p^T \\
V_p^{k,l}(W_q) =& 0 \qquad\qquad \text{if } q \neq p, p+1
\end{aligned} \tag{6.3.8}$$

where when $p \in \{0, N\}$, the only non-zero actions are $V_0^{k,l}(W_1)$ and $V_N^{k,l}(W_N)$.

The metric $g = \sum_{i=1}^N \mathrm{Tr}(\mathrm{d}W_i \mathrm{d}W_i^T)$ is expressed by a square matrix of dimension $d+(N+1)\frac{d(d-1)}{2}$. The matrix is block diagonalized in the following way: $g(l_j, V_p^{k,l}) = 0$ for all choices of index j, p, k, l, and $g(V_p^{k,l}, V_q^{m,n}) = 0$ as long as the pair k, l is not equal to the pair m, n. Therefore, g is block diagonalized as

$$g = \begin{pmatrix} g_l & & & \\ & g_V^{1,2} & & \\ & & \ddots & \\ & & & g_V^{d-1,d} \end{pmatrix} \tag{6.3.9}$$

with $(g_l)_{ij} := g(l_i, l_j)$ being $d \times d$ matrix, and $(g_V^{k,l})_{pq} := g(V_p^{k,l}, V_q^{k,l})$ being $(N+1) \times (N+1)$ matrix.

With further computation, we have g_l be equal to $N \times I_d$, a scalar matrix, and

$g_V^{k,l}$ being the following tridiagonal matrix

$$g_V^{k,l} = \begin{pmatrix} \frac{1}{2}(\lambda_k^2+\lambda_l^2) & -\lambda_k\lambda_l & & & & & \\ -\lambda_k\lambda_l & \lambda_k^2+\lambda_l^2 & -\lambda_k\lambda_l & & & & \\ & -\lambda_k\lambda_l & \lambda_k^2+\lambda_l^2 & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & -\lambda_k\lambda_l & \\ & & & & -\lambda_k\lambda_l & \lambda_k^2+\lambda_l^2 & -\lambda_k\lambda_l \\ & & & & & -\lambda_k\lambda_l & \frac{1}{2}(\lambda_k^2+\lambda_l^2) \end{pmatrix}_{(N+1)\times(N+1)} \tag{6.3.10}$$

By finding a proper matrix P such that $Pg_V^{k,l}P^T = I_{N+1}$ is diagonal, the vector fields defined by $U_p^{k,l} = \sum_{q=0}^{N} P_{pq}V_q^{k,l}$ for $p = 0, ..., N$ are orthonormal, i.e. $g(U_p^{k,l}, U_{p'}^{k,l}) = \delta_{pp'}$ for $p, p' = 0, ..., N$, therefore we are able to determine an orthonormal basis of $\mathfrak{X}(\mathcal{M})$.

The choice of P matrix is not unique. In the following we provide details about finding P associated to $U^{k,l}$ vector fields.

We first find eigenvalues and eigenvectors of the following submatrix of $g_V^{k,l}$ with the knowledge about Chebyshev polynomial.

$$g = \begin{pmatrix} \lambda_k^2+\lambda_l^2 & -\lambda_k\lambda_l & & & \\ -\lambda_k\lambda_l & \lambda_k^2+\lambda_l^2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \lambda_k^2+\lambda_l^2 & -\lambda_k\lambda_l \\ & & & -\lambda_k\lambda_l & \lambda_k^2+\lambda_l^2 \end{pmatrix}_{(N-1)\times(N-1)} . \tag{6.3.11}$$

Such g has decomposition $SgS^T = \Sigma$ with $SS^T = I_{N-1}$

$$\begin{aligned}
S_{pq} &= \sqrt{\frac{2}{N}}\sin(\frac{pq\pi}{N}) && p,q = 1,...,N-1\\
\sigma_p &= \lambda_k^2 + \lambda_l^2 - 2\lambda_k\lambda_l\cos(\frac{p\pi}{N}) && p = 1,...,N-1\\
\Sigma &= \mathrm{diag}(\sigma_1,...,\sigma_{N-1})\\
\det(g) &= \frac{\lambda_k^{2N} - \lambda_l^{2N}}{\lambda_k^2 - \lambda_l^2}
\end{aligned} \tag{6.3.12}$$

Now consider $g_V^{k,l}$, we still need other two vectors s_0, s_N to form an orthonormal basis. By the symmetry $(g_V^{k,l})_{i,j} = (g_V^{k,l})_{N-i,N-j}$ for $i,j = 0,...,N$, we can require s_0 and s_N to satisfy $(s_0)_i = (s_N)_{N-i}$ for $i = 0,...,N$.

As we require $s_0 g_V^{k,l}\tilde{S}^T = 0$, where $\tilde{S}$ is the extension of S to $(N+1)\times(N+1)$ matrix with $\tilde{S}_{pq} = \sqrt{\frac{2}{N}}\sin(\frac{pq\pi}{N})$ for $p,q = 0,...,N$, we have

$$s_0 g_V^{k,l} = \begin{pmatrix} c_0 & 0 & 0 & \cdots & 0 & c_N \end{pmatrix} \tag{6.3.13}$$

so s_0 is the linear combination of the first and last column of $(g_V^{k,l})^{-1}$.

Let $D_N = \det(g) = \frac{\lambda_k^{2N}-\lambda_l^{2N}}{\lambda_k^2-\lambda_l^2}$, using Cramer's rule and the tridiagonal shape of

$g_V^{k,l}$, it is not difficult to find

$$
\begin{aligned}
((g_V^{k,l})^{-1})_{0,i} =& \frac{1}{\det(g_V^{k,l})}(\lambda_k\lambda_l)^i \det(\begin{pmatrix} \lambda_k^2+\lambda_l^2 & -\lambda_k\lambda_l & & & \\ -\lambda_k\lambda_l & \lambda_k^2+\lambda_l^2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \lambda_k^2+\lambda_l^2 & -\lambda_k\lambda_l \\ & & & -\lambda_k\lambda_l & \frac{1}{2}(\lambda_k^2+\lambda_l^2) \end{pmatrix}_{(N-i)\times(N-i)}) \\
=& \frac{(\lambda_k\lambda_l)^i(D_{N-i+1}-\frac{1}{2}(\lambda_k^2+\lambda_l^2)D_{N-i})}{D_{N+2}-(\lambda_k^2+\lambda_l^2)D_{N+1}+\frac{1}{4}(\lambda_k^2+\lambda_l^2)^2D_N} \\
=& \frac{(\lambda_k\lambda_l)^i(\lambda_k^{2N-2i}+\lambda_l^{2N-2i})}{(\lambda_k^2-\lambda_l^2)(\lambda_k^{2N}-\lambda_l^{2N})} \\
\propto& \big((\frac{\lambda_k}{\lambda_l})^{N-i}+(\frac{\lambda_k}{\lambda_l})^{i-N}\big)\,, \qquad (6.3.14)
\end{aligned}
$$

then $((g_V^{k,l})^{-1})_{N,i} \propto \big((\frac{\lambda_k}{\lambda_l})^i + (\frac{\lambda_k}{\lambda_l})^{-i}\big)$ by symmetry.

Notice the following orthogonal property

$$
\sum_{p,q=0}^{N}(g_V^{k,l})_{pq}(\frac{\lambda_k}{\lambda_l})^p(\frac{\lambda_k}{\lambda_l})^{N-q} = 0\,, \qquad (6.3.15)
$$

we can choose $(s_0)_i = (\frac{\lambda_k}{\lambda_l})^i$ and $(s_N)_i = (\frac{\lambda_k}{\lambda_l})^{N-i}$, the corresponding norm are

$$
\begin{aligned}
\sigma_0 =& \sum_{p,q=0}^{N}(g_V^{k,l})_{pq}(\frac{\lambda_k}{\lambda_l})^p(\frac{\lambda_k}{\lambda_l})^q \\
=& \frac{1}{2}(\lambda_k^2-\lambda_l^2)((\frac{\lambda_k}{\lambda_l})^{2N}-1) \\
\sigma_N =& \frac{1}{2}(\lambda_l^2-\lambda_k^2)((\frac{\lambda_l}{\lambda_k})^{2N}-1) \qquad (6.3.16)
\end{aligned}
$$

Therefore, we found an orthogonal vectors of $g_V^{k,l}$. After normalization, we have

$$\begin{aligned}
P_{0j} =& \sqrt{\frac{2}{(\lambda_k^2-\lambda_l^2)(\lambda_k^{2N}-\lambda_l^{2N})}}\lambda_k^j\lambda_l^{N-j} \\
P_{ij} =& \sqrt{\frac{2}{N(\lambda_k^2+\lambda_l^2-2\lambda_k\lambda_l\cos(\frac{i\pi}{N}))}}\sin(\frac{ij\pi}{N}) \\
P_{Nj} =& \sqrt{\frac{2}{(\lambda_k^2-\lambda_l^2)(\lambda_k^{2N}-\lambda_l^{2N})}}\lambda_k^{N-j}\lambda_l^j \qquad\qquad i=1,...,N-1,\ j=0,...,N \\
&\text{with } Pg_V^{k,l}P^T = I_{N+1} \qquad (6.3.17)
\end{aligned}$$

□

Now we are going to consider the Riemannian submersion $\phi : \mathcal{M} \to \mathcal{N}$, $\phi(\underline{W}) = W_N W_{N-1} \cdots W_1$. $\mathcal{N}$ is the collection of matrices W with full rank, and the metric is determined through Riemannian submersion.

Theorem 36. The metric on $\mathcal{N}$

$$g_{\mathcal{N}} = \mathrm{Tr}(\mathcal{A}^{-1}(\mathrm{d}W), \mathrm{d}W^T) \qquad (6.3.18)$$

$$\mathcal{A}(Z) = \sum_{i=1}^{N}(WW^T)^{\frac{N-i}{N}}Z(W^TW)^{\frac{i-1}{N}} \qquad (6.3.19)$$

makes $\phi : \mathcal{M} \to \mathcal{N}$ a Riemannian submersion.

Remark 37. This metric is known in the work [7] from the analysis of gradient flow, by projection to the gradient flow of $\underline{W}_t$ into that of $W_t = \phi(\underline{W}_t)$. Here we provide a geometric view of this metric from Riemannian submersion.

Proof. To define $g_{\mathcal{N}}$ through Riemannian submersion, we should make $\phi_* : (\mathrm{Ker}\,\phi_*)^\perp \to T\mathcal{N}$ an isometry, where $\phi_* : T\mathcal{M} \to T\mathcal{N}$ is the differential of ϕ. Thus, finding an orthonormal basis of $(\mathrm{Ker}\,\phi_*)^\perp$ and determine the projection of the basis to $T\mathcal{N}$ actually provides an orthonormal basis of $T\mathcal{N}$, so $g_{\mathcal{N}}$ is obtained.

We claim that all $U_p^{k,l}$ vector fields with $1 \leq k < l \leq d$ and $p = 1, ..., N-1$ are in the kernel of ϕ_*. In fact, as each $U_p^{k,l}$ is a linear combination of vector fields $\{V_q^{k,l}\}_{q=1}^{N-1}$ and $V_q^{k,l}(W) = 0$, we conclude that $U_p^{k,l}(W) = 0$.

By counting dimension, $\operatorname{Ker}\phi_*$ has dimension $d+(N+1)\frac{d(d-1)}{2}-d^2 = (N-1)\frac{d(d-1)}{2}$, so $U_p^{k,l}$ in fact provides a basis of $\operatorname{Ker}\phi_*$. Therefore, the other vector fields gives an orthonormal basis of $(\operatorname{Ker}\phi_*)^\perp$, which are

$$L_j(W) =\sqrt{N}\lambda_j^{N-1}Q_N e_j e_j^T Q_0 \tag{6.3.20}$$

$$U_0^{k,l}(W) =\sqrt{\frac{\lambda_k^{2N}-\lambda_l^{2N}}{\lambda_k^2-\lambda_l^2}}Q_N e_l e_k^T Q_0 \tag{6.3.21}$$

$$U_N^{k,l}(W) =\sqrt{\frac{\lambda_k^{2N}-\lambda_l^{2N}}{\lambda_k^2-\lambda_l^2}}Q_N e_k e_l^T Q_0 \tag{6.3.22}$$

The orthonormal basis describes the metric $g_{\mathcal{N}}$ defined through Riemannian submersion completely. To confirm that this metric coincides with the one defined in 6.3.18, we write 6.3.18 into the following form: for $W = Q_N\Lambda^N Q_0$, defined the matrix-valued 1-form $\mathrm{d}M$ by $\mathcal{A}^{-1}(\mathrm{d}W) = Q_N \mathrm{d}M Q_0^T$, we have

$$\mathrm{d}W = \sum_{i=1}^{N} Q_N\Lambda^{2N-2i}\mathrm{d}M\Lambda^{2i-2}Q_0^T\,, \tag{6.3.23}$$

so

$$\begin{aligned}(Q_N^T \mathrm{d}W Q_0)_{jj} &=N\lambda_j^{2N-2}\mathrm{d}M_{jj}\,,\\ (Q_N^T \mathrm{d}W Q_0)_{kl} &=\frac{\lambda_k^{2N}-\lambda_l^{2N}}{\lambda_k^2-\lambda_l^2}\mathrm{d}M_{kl}\,.\end{aligned} \tag{6.3.24}$$

Now

$$
\begin{aligned}
g_{\mathcal{N}} =& \mathrm{Tr}(Q_N \mathrm{d}M Q_0^T (\mathrm{d}W)^T) \\
=& \mathrm{Tr}(\mathrm{d}M (Q_N^T \mathrm{d}W Q_0)^T) \\
=& \sum_{j=1}^{d} \frac{(Q_N^T \mathrm{d}W Q_0)_{jj}^2}{N\lambda_j^{2N-2}} + \sum_{1\le k<l\le d} \frac{\lambda_k^2-\lambda_l^2}{\lambda_k^{2N}-\lambda_l^{2N}} \big((Q_N^T \mathrm{d}W Q_0)_{kl}^2 + (Q_N^T \mathrm{d}W Q_0)_{lk}^2\big)
\end{aligned}
\tag{6.3.25}
$$

It is clear that the vector fields defined are orthonormal under $g_{\mathcal{N}}$, therefore $g_{\mathcal{N}}$ coincides with the metric obtained through Riemannian submersion. □

In the orthonormal basis described above, $\{V_p^{k,l} | 1 \le k < l \le d, 1 \le p \le N-1\}$ are vectors corresponding to the rotation in $(\mathrm{O}(d))^{N-1}$ that does not change the product $W = W_N W_{N-1} \cdots W_2 W_1$, which also gives a tangent space isomorphism between $T\phi^{-1}(W)$ and $(\mathrm{O}(d))^{N-1}$. With this understanding, we can evaluate the volume of fiber $\phi^{-1}(W)$ using change of variable rule.

6.4 Proof of main results

Proof of theorem 29. For $W = Q_N \Lambda^N Q_0^T$, consider the map $h : (\mathrm{O}(d))^{N-1} \to \phi^{-1}(W)$

$$h(Q_1, ..., Q_{N-1}) = (Q_N \Lambda Q_{N-1}^T, Q_{N-1} \Lambda Q_{N-2}^T, ..., Q_1 \Lambda Q_0^T) \tag{6.4.1}$$

It is clear that the push forward of the following vector fields

$$\alpha_p^{k,l}(Q_q) = \delta_{pq} Q_q \frac{-e_k e_l^T + e_l e_k^T}{\sqrt{2}} \tag{6.4.2}$$

for $1 \leq k < l \leq d$, $1 \leq p \leq N-1$ are just $V_p^{k,l}$, i.e. $h_*(\alpha_p^{k,l}) = V_p^{k,l}$.

Let the volume form corresponding to Haar measure on $(\mathrm{O}(d))^{N-1}$ be $\mathrm{d}v_{\mathrm{O}}$, the volume form corresponding to measure on fiber $\phi^{-1}(W)$ be $\mathrm{d}v_f$. The change of variable is written as

$$\mathrm{vol}(\phi^{-1}(W)) = \int_{\phi^{-1}(W)} \mathrm{d}v_f = \int_{(\mathrm{O}(d))^{N-1}} h^*(\mathrm{d}v_f) \tag{6.4.3}$$

where h^* is the pull back map of h.

Now we claim that $h^*(\mathrm{d}v_f) = \sqrt{\det(G)}\mathrm{d}v_{\mathrm{O}}$, where $G_{klp,mnq} := g(V_p^{k,l}, V_q^{m,n})$ is the metric matrix of dimension $(N-1)\frac{d(d-1)}{2}$. Let $\{\omega_p^{k,l} | 1 \leq k < l \leq d, 1 \leq p \leq N-1\}$ be the 1-form fields on $(\mathrm{O}(d))^{N-1}$ that is dual to vector fields $\{\alpha_p^{k,l} | 1 \leq k < l \leq d, 1 \leq p \leq N-1\}$, and $\{\Omega_p^{k,l} | 1 \leq k < l \leq d, 1 \leq p \leq N-1\}$ be the 1-form fields on $\phi^{-1}(W)$ that is dual to vector fields $\{V_p^{k,l} | 1 \leq k < l \leq d, 1 \leq p \leq N-1\}$, then the volume forms can be written as

$$
\begin{aligned}
\mathrm{d}v_{\mathrm{O}} =& \bigwedge_{1\leq k<l\leq d, 1\leq p\leq N-1} \omega_p^{k,l} \\
\mathrm{d}v_f =& \sqrt{\det(G)} \bigwedge_{1\leq k<l\leq d, 1\leq p\leq N-1} \Omega_p^{k,l}
\end{aligned} \tag{6.4.4}
$$

By $h_*(\alpha_p^{k,l}) = V_p^{k,l}$, we have the dual $\omega_p^{k,l} = h^*(\Omega_p^{k,l})$, so

$$
\begin{aligned}
h^*(\mathrm{d}v_f) =& \sqrt{\det(G)} h^*(\bigwedge_{1\leq k<l\leq d, 1\leq p\leq N-1} \Omega_p^{k,l}) \\
=& \sqrt{\det(G)} \bigwedge_{1\leq k<l\leq d, 1\leq p\leq N-1} \omega_p^{k,l} \\
=& \sqrt{\det(G)} \mathrm{d}v_{\mathrm{O}}
\end{aligned} \tag{6.4.5}
$$

As a result in theorem 34, the matrix G is block diagonalized with each block being 6.3.11, its determinant is given by

$$
\det(G) = \prod_{1\leq k<l\leq d} \frac{\lambda_k^{2N} - \lambda_l^{2N}}{\lambda_k^2 - \lambda_l^2} \tag{6.4.6}
$$

so we have

$$
\mathrm{vol}(\phi^{-1}(W)) = \prod_{1\leq k<l\leq d} \sqrt{\frac{\lambda_k^{2N} - \lambda_l^{2N}}{\lambda_k^2 - \lambda_l^2}} \int_{(\mathrm{O}(d))^{N-1}} \mathrm{d}v_{\mathrm{O}} = c_d^{N-1} \prod_{1\leq k<l\leq d} \sqrt{\frac{\lambda_k^{2N} - \lambda_l^{2N}}{\lambda_k^2 - \lambda_l^2}} \tag{6.4.7}
$$

Replacing λ_i by $\sigma_i^{\frac{1}{N}}$, we get the desired result. $\square$

With Riemannian submersion $\phi : \mathcal{M} \to \mathcal{N}$ determined, we are able to describe the $\mathcal{N}$-image of Brownian motion on $\mathcal{M}$. As the fiber $\phi^{-1}(W) \in \mathcal{M}$ is in fact the orbit of an isometry group $(\mathrm{O}(d))^{N-1}$, the Brownian motion on $\mathcal{M}$ projected to $\mathcal{N}$

is a Brownian motion with drift $\frac{1}{\beta}\text{grad}\log\text{vol}(\phi^{-1}(W))$. Similarly, for a process $\underline{W}_t$ solving Riemannian Langevin equation on $\mathcal{M}$ with energy function $E(W)$, the image $W_t = \phi(\underline{W}_t)$ solves Riemannian Langevin equation on $\mathcal{N}$ with free energy function $F(W) = E(W) - \frac{1}{\beta}S(W)$, where $S(W) = \log\text{vol}(\phi^{-1}(W))$.

Proof of theorem 31. The assertion of RLE 6.2.3 is standard result of RLE theory, and we only need to prove the explicit form of RLE 6.2.5.

For the term $\text{grad}_{g_\mathcal{N}} E(W)$, for each vector $Z \in T\mathcal{N}$, we have

$$g_\mathcal{N}(\text{grad}_{g_\mathcal{N}} E, Z)(W) = Z(E(W)) = \text{Tr}(\mathcal{D}E(W)(Z(W))^T) = g_\mathcal{N}(\mathcal{A}(\mathcal{D}E(W)), Z) \tag{6.4.8}$$

so

$$\text{grad}_{g_\mathcal{N}} E(W) = \mathcal{A}(\mathcal{D}E(W)) = \sum_{i=1}^{N} (W_t W_t^T)^{\frac{N-i}{N}} \mathcal{D}E(W_t)(W_t^T W_t)^{\frac{i-1}{N}} \tag{6.4.9}$$

The gradient of entropy term can be evaluated as follows, notice that the vector field $L_k = \frac{1}{\sqrt{N}}\frac{\partial}{\partial \lambda_k}$, and that $S(W) = \log\text{vol}(\phi^{-1}(W))$ is a function of $\{\lambda_k\}_{k=1}^d$

$$\begin{aligned}
\text{grad}_{g_\mathcal{N}} S(W) &= \sum_{k=1}^{d} g_\mathcal{N}(\text{grad}_{g_\mathcal{N}} S, L_k) L_k(W) \\
&= \sum_{k=1}^{d} L_k(S) L_k(W) \\
&= \sum_{k=1}^{d} \frac{1}{\sqrt{N}} \frac{\partial S}{\partial \lambda_k} L_k(W) \\
&= \sum_{k=1}^{d} \frac{\partial S}{\partial \lambda_k} Q_N \lambda_k^{N-1} e_k e_k^T Q_0^T
\end{aligned} \tag{6.4.10}$$

For the Brownian motion term, the diffusion term is given by the orthonormal basis $L_k(W)$ and $U_0^{k,l}(W), U_N^{k,l}(W)$, and the drift term is

$$\begin{aligned}\frac{1}{\beta}\Delta_{g_{\mathcal{N}}} W =& \frac{1}{\beta}\sum_{k=1}^{d}\Big(L_k(L_k(W)) - \nabla^{g_{\mathcal{N}}}_{L_k} L_k(W)\Big) \\ &+ \frac{1}{\beta}\sum_{1\le k<l\le d, p=0,N}\Big(U_p^{k,l}(U_p^{k,l}(W)) - \nabla^{g_{\mathcal{N}}}_{U_p^{k,l}} U_p^{k,l}(W)\Big)\end{aligned} \tag{6.4.11}$$

With some computation, we find that the second sum is 0, and that $\nabla^{g_{\mathcal{N}}}_{L_k} L_k = 0$, so

$$\begin{aligned}\frac{1}{\beta}\Delta_{g_{\mathcal{N}}} W =& \frac{1}{\beta}\sum_{k=1}^{d} L_k(L_k(W)) \\ =& \frac{1}{\beta}(N-1)\sum_{k=1}^{d} Q_N \lambda_k^{N-2} e_k e_k^T Q_0^T\end{aligned} \tag{6.4.12}$$

□

Chapter 7

Summary

This thesis contains the theory and several applications of Riemannian Langevin Equation in different problems. The most direct use of RLE to do Gibbs sampling problem is presented in chapters 3 4, and the stochastic relaxation of interior point method in conic programming is discussed in chapter 5. In chapters 2 6, the phenomena of entropy ascent companion with Brownian motion are emphasized. In chapter 2 it is used to generate random matrix models, and in 6 it is used to explain implicit regularization in deep linear network.

There are still many unknown questions and potential directions in each of chapters. For chapters 3 4, the convergence analysis of the Riemannian Langevin MCMC is not satisfied, and some criteria beyond Bakry-Émery curvature condition is expected. Also, the extension of RLE and Gibbs sampling to other manifolds is naturally motivated, e.g. the manifold of $n \times m$ matrices with fixed rank p.

In chapter 2, the parallel theorems in Dyson Brownian motion are not fully obtained. In particular, the proper scaling of the spectrum and the limit of spectrum measure is not solved. On the other hand, as a symplectic manifold, chapter 2 focused on its Riemannian structure, while the interaction between Brownian motion and symplectic manifold is also expected.

In chatper 5, a natural question about RLE in conic programming is the equilibrium measure. As the metric is generated by the barrier function which diverges at the boundary of cone, the geodesic distance from a interior point to a boundary point is infinite, and the volume of the manifold "concentrates" on the boundary. Thus RLE describes the evolution of a measure from the interior to the boundary, but the equilibrium measure on the boundary is not known.

The entropy ascent argument is used in chapter 6 to explain implicit regularization for DLN, and it is natural to ask about deep neutral networks. Without knowledge of balanced manifold, as well as the geometric structure in real deep neutral network, entropy argument cannot be applied directly, but the idea of maximizing volume of preimage in parameter space can be generalized.

In summary, Riemannian Langevin Equation lies in the interdisciplinary area of probability, Riemannian geometry and optimization, serves as a useful tool for practical problems, and also attracts theoretical interests. As a theory based on Riemannian geometry, RLE provides good explanation of entropy effect in optimization, but is also restricted by the compatible requirement of geometric structures. There are many prospective areas where RLE may be applied, many questions about RLE that need answers, and also many possible theories beyond RLE.

Bibliography

[1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds.* Princeton University Press, 2008.

[2] Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*, volume 118 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2010.

[3] Anton Arnold, Peter Markowich, Giuseppe Toscani, and Andreas Unterreiter. On logarithmic Sobolev inequalities, Csiszar-Kullback inequalities, and the rate of convergence to equilibrium for Fokker-Planck type equations. *Preprint*, 1998.

[4] Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[5] Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning*, pages 244–253. PMLR, 2018.

[6] Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Inf. Inference*, 11(1):307–353, 2022.

[7] Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Information and Inference: A Journal of the IMA*, 11(1):307–353, 2022.

[8] Jinho Baik, Percy Deift, and Toufic Suidan. *Combinatorics and random matrix theory*, volume 172 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2016.

[9] Dominique Bakry and Michel Émery. Diffusions hypercontractives. *Séminaire de probabilités de Strasbourg*, 19:177–206, 1985.

[10] Fabrice Baudoin. *An introduction to the geometry of stochastic flows*. World Scientific, 2004.

[11] Peter Baxendale and Theodore E Harris. Isotropic stochastic flows. *The Annals of Probability*, 14(4):1155–1179, 1986.

[12] DA Bayer and Jeffrey C Lagarias. Karmarkar's linear programming algorithm and Newton's method. *Mathematical Programming*, 50:291–330, 1991.

[13] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the bures–wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

[14] Vladimir I Bogachev, Nicolai V Krylov, Michael Röckner, and Stanislav V Shaposhnikov. *Fokker–Planck–Kolmogorov Equations*, volume 207. American Mathematical Society, 2015.

[15] C Le Bris and P-L Lions. Existence and uniqueness of solutions to Fokker–Planck type equations with irregular coefficients. *Communications in Partial Differential Equations*, 33(7):1272–1317, 2008.

[16] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[17] Marcus Brubaker, Mathieu Salzmann, and Raquel Urtasun. A Family of MCMC Methods on Implicitly Defined Manifolds. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 161–172, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

[18] Sébastien Bubeck and Ronen Eldan. The entropic barrier: a simple and optimal universal self-concordant barrier, 2015.

[19] Samuel Burer and Renato DC Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical programming*, 103(3):427–444, 2005.

[20] Simon Byrne and Mark Girolami. Geodesic Monte Carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40(4):825–845, 2013.

[21] Shiu-Yuen Cheng and Shing-Tung Yau. On the existence of a complete kähler metric on non-compact complex manifolds and the regularity of fefferman's equation. *Communications on Pure and Applied Mathematics*, 33(4):507–544, 1980.

[22] Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Conference on learning theory*, pages 300–323. PMLR, 2018.

[23] Xiang Cheng, Dong Yin, Peter Bartlett, and Michael Jordan. Stochastic gradient and Langevin processes. In *International Conference on Machine Learning*, pages 1810–1819. PMLR, 2020.

[24] Xiang Cheng, Jingzhao Zhang, and Suvrit Sra. Efficient sampling on riemannian manifolds via langevin mcmc. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 5995–6006. Curran Associates, Inc., 2022.

[25] Giovanni Ciccotti, Raymond Kapral, and Eric Vanden-Eijnden. Blue moon sampling, vectorial reaction coordinates, and unbiased constrained dynamics. *ChemPhysChem*, 6(9):1809–1814, 2005.

[26] Giovanni Ciccotti, Tony Lelievre, and Eric Vanden-Eijnden. Projection of diffusions on submanifolds: Application to mean force computation. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 61(3):371–408, 2008.

[27] Nadav Cohen, Govind Menon, and Zsolt Veraszto. Deep linear networks for matrix completion – an infinite depth limit, 2022.

[28] Nadav Cohen, Govind Menon, and Zsolt Veraszto. Deep linear networks for matrix completion—an infinite depth limit. *SIAM Journal on Applied Dynamical Systems*, 22(4):3208–3232, 2023.

[29] Richard WR Darling. *Differential forms and connections*. Cambridge University Press, 1994.

[30] Manfredo Perdigao Do Carmo and J Flaherty Francis. *Riemannian geometry*, volume 6. Springer, 1992.

[31] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32, 2019.

[32] Michel Émery. *Stochastic calculus in manifolds*. Universitext. Springer-Verlag, Berlin, 1989.

[33] Alessio Figalli. Existence and uniqueness of martingale solutions for SDEs with rough or degenerate coefficients. *Journal of Functional Analysis*, 254(1):109–153, 2008.

[34] Daniel JF Fox. A schwarz lemma for kähler affine metrics and the canonical potential of a proper convex cone. *Annali di Matematica Pura ed Applicata (1923-)*, 194(1):1–42, 2015.

[35] Khashayar Gatmiry and Santosh S Vempala. Convergence of the riemannian langevin algorithm. *arXiv preprint arXiv:2204.10818*, 2022.

[36] Rong Ge, Holden Lee, Jianfeng Lu, and Andrej Risteski. Efficient sampling from the bingham distribution. In *Algorithmic Learning Theory*, pages 673–685. PMLR, 2021.

[37] Charles J Geyer. Markov chain monte carlo lecture notes. *Course notes, Spring Quarter*, 80, 1998.

[38] 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, 2011.

[39] Osman Güler. Barrier functions in interior point methods. *Mathematics of Operations Research*, 21(4):860–885, 1996.

[40] Roland Hildebrand. Canonical barriers on convex cones. *Mathematics of operations research*, 39(3):841–850, 2014.

[41] Roland Hildebrand. *Conic optimization: affine geometry of self-concordant barriers and copositive cones*. Habilitation à diriger des recherches, Université Grenoble Alpes, July 2017.

[42] Richard Holley and Daniel W. Stroock. Logarithmic sobolev inequalities and stochastic ising models. *Journal of Statistical Physics*, 46:1159–1194, 1987.

[43] Elton P. Hsu. *Stochastic analysis on manifolds*, volume 38 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2002.

[44] Ching-Peng Huang, Dominik Inauen, and Govind Menon. Motion by mean curvature and Dyson Brownian motion. *Electron. Commun. Probab.*, 28:1–10, 2023.

[45] Jiaoyang Huang and Colin McSwiggen. Asymptotics of generalized Bessel functions and weight multiplicities via large deviations of radial Dunkl processes, 2023.

[46] Wen Huang, Kyle A. Gallivan, and Xiangxiong Zhang. Solving PhaseLift by low-rank Riemannian optimization methods for complex semidefinite constraints. *SIAM Journal on Scientific Computing*, 39(5):B840–B859, 2017.

[47] Nobuyuki Ikeda and Shinzo Watanabe. *Stochastic differential equations and diffusion processes*, volume 24 of *North-Holland Mathematical Library*. North-Holland Publishing Co., Amsterdam; Kodansha, Ltd., Tokyo, second edition, 1989.

[48] Dominik Inauen and Govind Menon. Stochastic nash evolution, 2023.

[49] Pauwels JE. Riemannian submersions of brownian motions. *Stochastics: An International Journal of Probability and Stochastic Processes*, 29(4):425–436, 1990.

[50] Pauwels J.E. Riemannian submersions of brownian motions. *Stochastics and Stochastic Reports*, 29(4):425–436, 1990.

[51] Galin L. Jones. On the Markov chain central limit theorem. *Probability Surveys*, 1(none):299 – 320, 2004.

[52] Jürgen Jost. *Riemannian geometry and geometric analysis*. Springer, 2011.

[53] Michel Journée, Francis Bach, P-A Absil, and Rodolphe Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 20(5):2327–2351, 2010.

[54] Narendra Karmarkar. A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 302–311, 1984.

[55] Wilhelm Klingenberg. *Riemannian geometry*, volume 1. Walter de Gruyter, 1995.

[56] Peter E. Kloeden and Eckhard Platen. *Numerical solution of stochastic differential equations*, volume 23 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1992.

[57] Hiroshi Kunita. *Stochastic flows and stochastic differential equations*, volume 24. Cambridge university press, 1990.

[58] Jonathan Leake, Colin McSwiggen, and Nisheeth K Vishnoi. Sampling matrices from harish-chandra–itzykson–zuber densities with applications to quantum inference and differential privacy. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1384–1397, 2021.

[59] Michel Ledoux. Logarithmic sobolev inequalities for unbounded spin systems revisited. *Séminaire de Probabilités XXXV*, pages 167–194, 2001.

[60] John M. Lee. *Introduction to smooth manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, New York, second edition, 2013.

[61] J. T. Lewis. Brownian motion on a submanifold of Euclidean space. *Bull. London Math. Soc.*, 18(6):616–620, 1986.

[62] Mufan Li and Murat A Erdogdu. Riemannian langevin algorithm for solving semidefinite programs. *Bernoulli*, 29(4):3093–3113, 2023.

[63] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019.

[64] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[65] Jun S Liu and Jun S Liu. *Monte Carlo strategies in scientific computing*, volume 75. Springer, 2001.

[66] Paul Malliavin. *Stochastic analysis*, volume 313 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1997.

[67] Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research*, 18, 2017.

[68] P. A. Markowich and C. Villani. On the trend to equilibrium for the Fokker-Planck equation: an interplay between physics and functional analysis. volume 19, pages 1–29. 2000. VI Workshop on Partial Differential Equations, Part II (Rio de Janeiro, 1999).

[69] Estelle Massart and P-A Absil. Quotient geometry with simple geodesics for the manifold of fixed-rank positive-semidefinite matrices. *SIAM Journal on Matrix Analysis and Applications*, 41(1):171–198, 2020.

[70] Estelle Massart, Julien M Hendrickx, and P-A Absil. Curvature of the manifold of fixed-rank positive-semidefinite matrices endowed with the bures–wasserstein metric. In *Geometric Science of Information: 4th International Conference, GSI 2019, Toulouse, France, August 27–29, 2019, Proceedings*, pages 739–748. Springer, 2019.

[71] Govind Menon and Tianmin Yu. The riemannian langevin equation and conic programs, 2023.

[72] Govind Menon and Tianmin Yu. Siegel brownian motion, 2023.

[73] Gilles Meyer, Silvère Bonnabel, and Rodolphe Sepulchre. Regression on fixed-rank positive semidefinite matrices: a Riemannian approach. *The Journal of Machine Learning Research*, 12:593–625, 2011.

[74] Ankur Moitra and Andrej Risteski. Fast convergence for langevin with matrix manifold structure. In *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, 2020.

[75] Andrea Montanari. Sampling, diffusions, and stochastic localization. *arXiv preprint arXiv:2305.10690*, 2023.

[76] Barrett O'Neill. The fundamental equations of a submersion. *Michigan Mathematical Journal*, 13(4):459–469, 1966.

[77] Tommaso Pacini. Mean curvature flow, orbits, moment maps. *Transactions of the American Mathematical Society*, 355(8):3343–3357, 2003.

[78] Tommaso Pacini. Mean curvature flow, orbits, moment maps. *Trans. Amer. Math. Soc.*, 355(8):3343–3357, 2003.

[79] Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Homogenization of sgd in high-dimensions: Exact dynamics and generalization

properties, 2022.

[80] F. W. PONTING and H. S. A. POTTER. THE VOLUME OF ORTHOGONAL AND UNITARY SPACE. *The Quarterly Journal of Mathematics*, os-20(1):146–154, 01 1949.

[81] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951.

[82] L Chris G Rogers and David Williams. *Diffusions, Markov processes and martingales: Volume 2, Itô calculus*, volume 2. Cambridge university press, 2000.

[83] Carl Ludwig Siegel. Symplectic geometry. *Amer. J. Math.*, 65:1–86, 1943.

[84] Daniel W Stroock. *An introduction to the analysis of paths on a Riemannian manifold*. Number 74. American Mathematical Soc., 2000.

[85] Daniel W Stroock. Brownian motion on a submanifold. *Communications in analysis and geometry*, 12(1):165–182, 2004.

[86] Daniel W Stroock and SR Srinivasa Varadhan. *Multidimensional diffusion processes*, volume 233. Springer Science & Business Media, 1997.

[87] Teiji Takagi. On an algebraic problem related to an analytic theorem of Carathéodory and Fejér and on an allied theorem of Landau. *Japanese Journal of Mathematics: transactions and abstracts*, 1:83–93, 1924.

[88] Abiy Tasissa and Rongjie Lai. Exact reconstruction of euclidean distance geometry problem using low-rank matrix completion. *IEEE Transactions on Information Theory*, 65(5):3124–3144, 2018.

[89] Guiseppe Toscani and Cédric Villani. On the trend to equilibrium for some dissipative systems with slowly increasing a priori bounds. *Journal of Statistical Physics*, 98:1279–1309, 2000.

[90] M Van Den Berg and JT806240 Lewis. Brownian motion on a hypersurface. *Bulletin of the London Mathematical Society*, 17(2):144–150, 1985.

[91] Bart Vandereycken, P-A Absil, and Stefan Vandewalle. Embedded geometry of the set of symmetric positive semidefinite matrices of fixed rank. In *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*, pages 389–392. IEEE, 2009.

[92] Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of langevin dynamics based algorithms for nonconvex optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[93] Tianmin Yu, Shixin Zheng, Jianfeng Lu, Govind Menon, and Xiangxiong Zhang. Riemannian Langevin equations for PSD matrices of fixed rank, 2023.

[94] Tianmin Yu, Shixin Zheng, Jianfeng Lu, Govind Menon, and Xiangxiong Zhang. Riemannian Langevin Monte Carlo schemes for sampling PSD matrices with fixed rank. *arXiv preprint arXiv:2309.04072*, 2023.

[95] Emilio Zappa, Miranda Holmes-Cerfon, and Jonathan Goodman. Monte Carlo on manifolds: sampling densities and integrating functions. *Communications on Pure and Applied Mathematics*, 71(12):2609–2647, 2018.

[96] Shixin Zheng, Wen Huang, Bart Vandereycken, and Xiangxiong Zhang. Riemannian optimization using three different metrics for Hermitian PSD fixed-rank constraints: an extended version. *arXiv:2204.07830*, 2022.