Fast Convergence of Langevin Dynamics on Manifold: Geodesics meet Log-Sobolev

Qi Lei
ODEN
UT Austin
leiqi@ices.utexas.edu

Ioannis Panageas
ISTD
SUTD
ioannis@sutd.edu.sg

Xiao Wang
ISTD
SUTD
xiao_wang@sutd.edu.sg

Abstract

Sampling is a fundamental and arguably very important task with numerous applications in Machine Learning. One approach to sample from a high dimensional distribution e^{-f} for some function f is the Langevin Algorithm (LA). Recently, there has been a lot of progress in showing fast convergence of LA even in cases where f is non-convex, notably Vempala and Wibisono [2019], Moitra and Risteski [2020] in which the former paper focuses on functions f defined in $\mathbb{R}^n$ and the latter paper focuses on functions with symmetries (like matrix completion type objectives) with manifold structure. Our work generalizes the results of Vempala and Wibisono [2019] where f is defined on a manifold M rather than $\mathbb{R}^n$. From technical point of view, we show that KL decreases in a geometric rate whenever the distribution e^{-f} satisfies a log-Sobolev inequality on M.

1 Introduction

We focus on the problem of sampling from a distribution $e^{-f(x)}$ supported on a Riemannian manifold M with standard volume measure. Sampling is a fundamental and arguably very important task with numerous applications in machine learning and Langevin dynamics is a quite standard approach. There is a growing interest in Langevin algorithms, e.g. Welling and Teh [2011], Wibisono [2018], Dalalyan [2017a], due to its simple structure and the good empirical behavior. The classic Riemannian Langevin algorithm, e.g. Girolami and Calderhead [2011], Patterson and Teh [2013], Zhang et al. [2020], is used to sample from distributions supported on $\mathbb{R}^n$ (or a subset D) by endowing $\mathbb{R}^n$ (or D) a Riemannian structure. Beyond the classic application of Riemannian Langevin Algorithm (RLA), recent progress in Domingo-Enrich et al. [2020], Moitra and Risteski [2020] shows that sampling from a distribution on a manifold has application in matrix factorization, principal component analysis, matrix completion, mean field and continuous games and GANs. Formally, a game with finite number of agents is called continuous if the strategy spaces are continuous, either a finite dimensional differential manifold or an infinite dimensional Banach manifold Ratliff et al. [2013, 2016], Domingo-Enrich et al. [2020]. The mixed strategy is then a probability distribution on the strategy manifold and mixed Nash equilibria can be approximated by Langevin dynamics.

Geodesic Langevin Algorithm (GLA). In order to sample from a distribution on M, geodesic based algorithms (e.g. Geodesic Monte Carlo and Geodesic MCMC) are considered in Byrne and Girolami [2013], Liu et al. [2016], where a geodesic integrator is used in the implementation. We propose a Geodesic Langevin Algorithm (GLA) as a natural generalization of unadjusted Langevin algorithm (ULA) from the Euclidean space to manifold M. The benefit of GLA is to leverage sufficiently the geometric information (curvature, geodesic distance, isoperimetry) of M while keeping the structure of the algorithm simple enough, so that we can obtain a non-asymptotic convergence guarantee of the algorithm. In local coordinate systems, the Riemannian metric is represented by a matrix $g = \{g_{ij}\}$, see Definition 3.3. We denote g^{ij} the ij-th entry of the inverse

matrix g^{-1} of g, and $|g| = \det(g_{ij})$, the determinant of the matrix $\{g_{ij}\}$. Then GLA is the stochastic process on M that is defined by

$$x_{k+1} = \text{Exp}_{x_k}(\epsilon F + \sqrt{2\epsilon g^{-1}}\xi_0) \tag{1}$$

where $F = (F_1, ..., F_n)$ with

$$F_i = -\sum_j g^{ij}\frac{\partial f}{\partial x_j} + \frac{1}{\sqrt{|g|}}\sum_j \frac{\partial}{\partial x_j}\left(\sqrt{|g|}g^{ij}\right), \tag{2}$$

$\epsilon > 0$ is the stepsize, $\xi_0 \sim \mathcal{N}(0, I)$ is the standard Gaussian noise, and $\text{Exp}_x(\cdot) : T_xM \to M$ is the exponential map (Definition 3.5). Clearly GLA is a two-step discretization scheme of the Riemannian Langevin equation

$$dX_t = F(X_t)dt + \sqrt{2g^{-1}}dB_t$$

where F is given by (2). Suppose the position at time k is x_k, then the next position x_{k+1} can be obtained by the following tangent-geodesic composition:

1. Tangent step: Take a local coordinate chart $\varphi : U_{x_k} \to \mathbb{R}^n$ at x_k, this map induces the expression of g_{ij} and g^{ij}, then compute the vector $v = \epsilon F + \sqrt{2\epsilon g^{-1}}\xi_0$ in tangent space $T_{x_k}M$;

2. Geodesic step: Solve the geodesic equation (a second order ODE) whose solution is a curve $\gamma(t) \subset \varphi(U_{x_k})$, such that the initial conditions satisfy $\gamma(0) = \varphi(x_k)$ and $\gamma'(0) = v$. Then let $x_{k+1} = \gamma(1)$ be the updated point.

The exponential map and ODE solver for geodesic equations is commonly used in sampling algorithms on manifold, e.g. Vempala and Lee [2017], Byrne and Girolami [2013], Liu et al. [2016]. We will discuss on other approximations of the exponential map without solving ODEs through illustrations in a later section. Figure 1 gives an intuition of GLA on the unit sphere where the exponential map is $\text{Exp}_x(v) = \cos(\|v\|)x + \sin(\|v\|)\frac{v}{\|v\|}$.
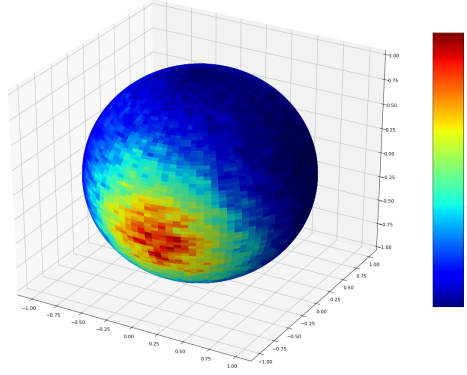

Figure 1: $f(x_1, x_2, x_3) = x_1^2 + 3.05x_2^2 - 0.9x_3^2 + 1.1x_1x_2 + -1.02x_2x_3 + 2.1x_3x_1$, $\epsilon = 0.1$, Iterations: 100k.

The main result on convergence is stated as follows.

Theorem 1.1 (Informal). *Let M be a closed n-dimensional manifold (Definition 3.2). Suppose that $\nu = e^{-f(x)}$ is a distribution on M with $\alpha > 0$ the log-Sobolev constant. Then there exists a real number $C > 0$, such that by choosing stepsize ϵ properly based on the Lipschitz constant of the Riemannian gradient of f, log-Sobolev constant of the target distribution ν, dimension and curvature of M, the KL divergence $H(\rho_k|\nu)$ decreases along the GLA iterations rapidly in the sense that*

$$H(\rho_k|\nu) \leq e^{-\frac{k}{2}\alpha\epsilon}H(\rho_0|\nu) + \frac{C\epsilon nL^2}{\alpha}.$$

The same as unadjusted Langevin algorithm (ULA) in Euclidean space, GLA is a biased algorithm that converges to a distribution different from $e^{-f(x)}$. But from the above theorem, one can set up the error $\delta > 0$ and take stepsize $\epsilon < \frac{\alpha\delta}{2CnL^2}$ while satisfying the standard from above theorem, and GLA will reach the error $H(\rho_k|\nu) < \delta$ after $k > \frac{2}{\alpha\epsilon}\log\frac{2H(\rho_0|\nu)}{\delta}$ iterations. Practically we need a lower bound estimate for α. With additional condition on Ricci curvature, this lower bound can be chosen based on the diameter of M by Theorem 3.13.

Our main technical contributions are:

- A non-asymptotic convergence guarantee for Geodesic Langevin algorithm on closed manifold is provided, with the help of log-Sobolev inequality.
- The framework of this paper serves as the first step understanding to the rate of convergence in sampling from distributions on manifold with log-Sobolev inequality, and can be generalized to prove non-asymptotic convergence results for more general settings and more subtle algorithms, i.e., for open manifolds and unbiased algorithms.

Comparison to literarture The typical difference between algorithm (1) and the classic RLA is the use of exponential map. As $\epsilon \to 0$, both GLA and RLA boil down to the same continuous time Langevin equation in the local coordinate system:

$$dX_t = F(X_t)dt + \sqrt{2g^{-1}}dB_t$$

where F is given by (2) and B_t is the standard Brownian motion in $\mathbb{R}^n$. The direct Euler-Maryuyama discretization iterates in the way that $x_{k+1} = x_t + \epsilon F(x_k) + \sqrt{2\epsilon g^{-1}(x_k)}\xi_0$. However, by adding the vector $\epsilon F(x_k) + \sqrt{2\epsilon g^{-1}(x_k)}\xi_0$ that is in the tangent space to a point x_k that is on the manifold M has no intrinsic geometric meaning, since the resulted point x_{k+1} is indeed not in M. The exponential map just gives a way to pull x_{k+1} back to M. On the other hand, since RLA is firstly used to sample from distributions on $\mathbb{R}^n$ (or its domain) with a Riemannian structure, Roberts and Stramer [2002], Girolami and Calderhead [2011], Patterson and Teh [2013], Smith et al. [2018], this requires a global coordinate system of M, i.e. M is covered by a single coordinate chart and the iterations do not transit between different charts. But this makes it difficult to use RLA when there are inevitably multiple coordinate charts on M. More sophisticated algorithms like Geodesic MCMC Liu et al. [2016] is used to transit between different coordinate charts, but to the best knowledge of the authors, the rate of convergence is missing in the literature.

2   Related Works

Unadjusted Langevin algorithm (ULA) when sampling from a strongly logconcave density in Euclidean space has been studied extensively in the literature. The bounds for ULA is known in Cheng and Bartlett [2018], Dalalyan [2017a], Dalalyan and Karagulyan [2019], Durmus et al. [2018]. The case when f is strongly convex and has Lipschitz gradient is studied by Dalalyan [2017b], Durmus and Moulines [2017], Durmus and Mounline [2019]. Since ULA is biased because of the discretization, i.e. it converges to a limit distribution that is different from that from continuous Langevin equation. the Metropolis-Hastings correction is widely used to correct this bias, e.g. Roberts and Tweedie [1996], Dwivedi et al. [2018]. A simplified correction algorithm is proposed by Wibisono [2018] that is called symmetrized Langevin algorithm with a smaller bias than ULA. Convergence results is obtained for Proximal Langevin algorithm (PLA) in Wibisono [2019]. In the case where the target distribution is log-concave, there are other algorithms proven to converge rapidly, i.e., Langevin Monte Carlo by Bernton [2018], ball walk and hit-and-run Kannan et al. [1997], Lovász and Vempala [2006a, 2007, 2006b], and Hamiltonian Monte Carlo by Durmus et al. [2017], Vempala and Lee [2018], Mangoubi and Vishnoi [2018]. The underdamped version of the Langevin dynamics under log-Sobolev inequality is studied by Ma et al. [2019], where an iteration complexity for the discrete time algorithm that has better dependence on the dimension is provided. A coupling approach is used by Eberle et al. [2018] to quantify convergence to equilibrium for Langevin dynamics that yields contractions in a particular Wasserstein distance and provides precise bounds for convergence to equilibrium. The case where the densities that are neither smooth nor log-concave is studied in Luu et al. [2017] and asymptotic consistency guarantees is provided. For the Wasserstein distance, Cheng et al. [2018], Majka et al. [2018], Raginsky et al. [2017] provide convergence bound. An earlier research on stochastic gradient Langevin dynamics with application in Bayesian learning is proposed

by Welling and Teh [2011], The Langevin Monte Carlo with a weaker smoothness assumption is studied by Chatterji et al. [2019]. In order to improve sample quality, Gorham and Mackey [2017] develops a theory of weak convergence for kernel Stein discrepancy based on Stein's method. In general, sampling from non log-concave densities is hard, Ge et al. [2018] gives an exponential lower bound on the number of queries required.

The Riemannian Langevin algorithm has been studied in different extent. Related to volume computation of a convex body in Euclidean space, one can endow the interior of a convex body the structure of a Hessian manifold and run geodesic (with respect to the Hessian metric) random walk Vempala and Lee [2017] that is a discretization scheme of a stochastic process with uniform measure as the stationary distribution. The rigorous proof of the convergence of Riemannian Hamiltonian Monte Carlo for sampling Gibbs distribution and uniform distribution in a polytope is given by Vempala and Lee [2018]. In sampling non-uniform distribution, Zhang et al. [2020] gives a discretization scheme related to mirror descent and a non-asymptotic upper bound on the sampling error of the Riemannian Langevin Monte Carlo algorithm in Hessian manifold. The mirrored Langevin is firstly considered by Hsieh et al. [2018] and a non-asymptotic rate is obtained and generalized to the case when only stochastic gradients (mini-batch) are available. An affine invariant perspective of continuous time Langevin dynamics for Bayesian inference is studied in Inigo et al. [2019]. Positive curvature is used to show concentration results for Hamiltonian Monte Carlo in Seiler et al. [2014].

3 Preliminaries

For a complete introduction to Riemannian manifold and stochastic analysis on manifold, we recommend Lee [2018] and Hsu [2002] for references.

3.1 Riemannian geometry

Definition 3.1 (Manifold). A C^k-differentiable, n-dimensional manifold is a topological space M, together with a collection of coordinate charts $\{(U_\alpha, \varphi_\alpha)\}$, where each φ_α is a C^k-diffeomorphism from an open subset $U_\alpha \subset M$ to $\mathbb{R}^n$. The charts are compatible in the sense that, whenever $U_\alpha \cap U_\beta \neq \emptyset$, the transition map $\varphi_\alpha \circ \varphi_\beta^{-1}(U_\beta \cap U_\alpha) \to \mathbb{R}^n$ is of C^k.

Definition 3.2 (Closed manifold). A manifold M is called *closed* if M is compact and has no boundary.

Typical examples of closed manifolds include sphere and torus.

Definition 3.3 (Riemannian metric). A Riemannian manifold (M, g) is a differentiable manifold M with a Riemannian metric g defined as the inner product on the tangent space $T_x M$ for each point x, $g(\cdot, \cdot) : T_x M \times T_x M \to \mathbb{R}$. Then length of a smooth path $\gamma : [0, 1] \to M$ is $|\gamma| = \int_0^1 \sqrt{g(\gamma'(t), \gamma'(t))} dt$. In a local coordinate chart, g is represented by a $n \times n$ symmetric positive definite matrix with entries g_{ij}.

Definition 3.4 (Geodesic). We call a curve $\gamma(t) : [0, 1] \to M$ a geodesic if it satisfies both of the following conditions:

1. The curve $\gamma(t)$ is parametrized with constant speed, i.e. $\left\| \frac{d}{dt} \gamma(t) \right\|_{\gamma(t)}$ is constant for $t \in [0, 1]$.

2. The curve is the locally shortest length curve between $\gamma(0)$ and $\gamma(1)$, i.e. for any family of curve $c(t, s)$ with $c(t, 0) = \gamma(t)$ and $c(0, s) = \gamma(0)$ and $c(1, s) = \gamma(1)$, we have $\frac{d}{ds}|_{s=0} \int_0^1 \left\| \frac{d}{dt} c(t, s) \right\|_{c(t,s)} dt = 0$.

We use $\gamma_{x \to y}$ to denote the geodesic from x to y ($\gamma_{x \to y}(0) = x$ and $\gamma_{x \to y}(1) = y$). The most important property of a geodesic $\gamma(t)$ is that the time derivative $\dot{\gamma}(t)$ as a vector field, has 0 covariant derivative, i.e. $\nabla_{\dot{\gamma}(t)} \dot{\gamma}(t) = 0$. This property boils down to a second order ODE in local coordinate systems,

$$\ddot{\gamma}_i(t) + \sum_{j,k} \Gamma_{jk}^i (\gamma(t)) \dot{\gamma}_j(t) \dot{\gamma}_k(t) = 0$$

for $i \in [n]$, where Γ^i_{jk} are the Christoffel symbols. Given a initial position $\gamma(0)$ and initial velocity $\dot{\gamma}(0)$, by the fundamental theorem of ODE, there exists a unique solution satisfying the geodesic equation. This is the principle we can use the ODE solver in GLA.

Definition 3.5 (Exponential map). The exponential map $\mathrm{Exp}_x(v)$ is maps $v \in T_x M$ to $y \in M$ such that there exists a geodesic γ with $\gamma(0) = x$, $\gamma(1) = y$ and $\gamma'(0) = v$.

The exponential map can be thought of moving a point along a vector in manifold in the sense that the exponential map in $\mathbb{R}^n$ is nothing but $\mathrm{Exp}_x(v) = x + v$. The exponential map on sphere at x with direction v is $\mathrm{Exp}_x(v) = \cos(\|v\|)x + \sin(\|v\|)\frac{v}{\|v\|}$.

Definition 3.6 (Parallel transport). The parallel transport Γ^y_x is a map that transport $v \in T_x M$ to $\Gamma^y_x \in T_x M$ along $\gamma_{x \to y}$ such that the vector stays constant by satisfying a zero-acceleration condition.

Next, we refer the definition of Riemannian gradient and divergence only in local coordinate systems that is used in this paper.

Definition 3.7 (Gradient and Divergence). In local coordinate system, the gradient of f and the divergence of a vector field $V = \sum_i V_i \frac{\partial}{\partial x_i}$ on a Riemannian manifold is given by

$$\mathrm{grad} f = \sum_{i,j} g^{ij} \frac{\partial f}{\partial x_i} \frac{\partial}{\partial x_i} \quad \text{and} \quad \mathrm{div} V = \frac{1}{\sqrt{|g|}} \sum_i \frac{\partial}{\partial x_i} \left(\sqrt{|g|} V_i \right)$$

where g^{ij} is the ij-th entry of the inverse matrix g^{-1} of g, $|g| = \det(g_{ij})$.

Definition 3.8 (Lipschitz gradient). f is of Lipschitz gradient if there exists a constant $L > 0$ such that
$$\|\mathrm{grad} f(y) - \Gamma^y_x \mathrm{grad} f(x)\| \leq L d(x, y) \quad \text{for all} \quad x, y \in M$$
where $d(x, y)$ is the geodesic distance between x and y, and Γ^y_x is the parallel transport from x to y, see Definition 3.6.

3.2 Stochastic differential equations

Let $\{X_t\}_{t \geq 0}$ be a stochastic process in $\mathbb{R}^n$ and B_t be the standard Brownian motion in $\mathbb{R}^n$.

Fokker-Planck Equation For any stochastic differential equation of the form
$$dX_t = F(X_t, t)dt + \sigma(X_t, t)dB_t,$$

the probability density of the SDE is given by the PDE

$$\frac{\partial \rho(x, t)}{\partial t} = -\sum_{i=1}^n \frac{\partial}{\partial x_i} \left(F_i(x, t) \rho(x, t) \right) + \sum_{i=1}^n \sum_{j=1}^n \frac{\partial^2}{\partial x_i \partial x_j} \left(A_{ij}(x, t) \rho(x, t) \right)$$

where $A = \frac{1}{2} \sigma \sigma^\top$, i.e. $A_{ij} = \frac{1}{2} \sum_{k=1}^n \sigma_{ik}(x, t) \sigma_{jk}(x, t)$

3.3 Distributions on manifold

Let ρ and ν be probability distribution on M that is absolutely continuous with respect to the Riemannian volume measure (denoted by dx) on M.

Definition 3.9 (KL divergence). The Kullback-Leibler (KL) divergence of ρ with respect to ν is
$$H(\rho|\nu) = \int_M \rho(x) \log \frac{\rho(x)}{\nu(x)} dx.$$

Definition 3.10 (Wasserstein distance). The Wasserstein distance between μ and ν is defined to be
$$W_2(\mu, \nu) = \inf\{\sqrt{\mathbb{E}[d(X, Y)^2]} : \mathrm{law}(X) = \mu, \mathrm{law}(Y) = \nu\}.$$

Definition 3.11 (Talagrand inequality). The probability measure ν satisfies a Talagrand inequality with constant $\alpha > 0$ if for all probability measure ρ, absolutely continuous with respect to ν, with finite moments of order 2,
$$W_2(\rho, \nu)^2 \leq \frac{2}{\alpha} H(\rho|\nu)$$

Definition 3.12 (Log-Sobolev inequality). A probability measure ν on M is called to satisfy the logarithmic Sobolev inequality (LSI) if there exists a constant $\alpha > 0$ such that

$$\int_M g^2 \log g^2 d\nu - \left(\int_M g^2 d\nu \right) \log \left(\int_M g^2 d\nu \right) \leq \frac{2}{\alpha} \int_M \|\text{grad}g\|^2 d\nu,$$

for all smooth functions $g : M \to \mathbb{R}$ with $\int_M g^2 \leq \infty$. The largest possible constant α is called the logarithmic Sobolev constant (LSC).

Estimate of the log-Sobolev constant It is well known that a compact manifold always satisfies log-Sobolev inequality Gross [1993], Otto and Villani [2000], Rothaus [1981, 1986], Wang [1997a,b]. Practically we need a specific lower bound α_0 of the log-Sobolev constant α, so that we can choose stepsize $\epsilon \leq \frac{\alpha_0 \delta}{2CnL^2} \leq \frac{\alpha \delta}{2CnL^2}$ for a given error bound δ for the KL divergence, see the discussion next to Theorem 1.1. The estimate of α_0 is closely related to the Ricci curvature of M and the first eigenvalue λ_1 of the Laplace-Beltrami operator on M, see the definitions in Appendix. For the case where M is compact and the Ricci curvature is non-negative, the lower bound estimate for α is clear from the following theorem.

Theorem 3.13 (Theorem 7.3, Ledoux [1999]). *Let M be a compact Riemannian manifold with diameter D and non-negative Ricci curvature. Then the log-Sobolev constant α satisfies $\alpha \geq \frac{\lambda_1}{1+2D\sqrt{\lambda_1}}$. In particular, $\alpha \geq \frac{\pi^2}{(1+2\pi)D^2}$.*

The last inequality shows that we can choose a lower bound $\alpha_0 = \frac{\pi^2}{(1+2\pi)D^2}$ that only depends on the diameter of M. For more results on estimate of the log-Sobolev constant, we refer to Wang [1997a,b].

4 Main Results

4.1 Technical Overview

Wasserstein gradient flow. The equivalence between Langevin dynamics and optimization in the space of densities is based on the result of Jordan et al. [1998], Wibisono [2018] that the Langevin dynamics captures the gradient flow of the relative entropy functional in the space of densities with the Wasserstein metric. As a result, running the Langevin dynamics is equivalent to sampling from the stationary distribution of the Wasserstein gradient flow asymptotically. To minimize $\int_M f(x)\rho(x)dx$ with respect to $\rho \in \mathcal{P}(M)$, we consider the entropy regularized functional of ρ defined as follows,

$$\mathcal{L}(\rho) = \mathcal{F}(\rho) + H(\rho)$$

where $\mathcal{F}(\rho) = \int_M f(x)\rho(x)dx$ and $H(\rho) = \int_M \rho(x) \log \rho(x)dx$ that is the negative Shannon entropy $h(\rho) = -\int_M \rho(x) \log \rho(x)dx$. According to Santanmbrogio [2015], the Wasserstein gradient flow associated with functional $\mathcal{L}$ is the Fokker-Planck equation

$$\frac{\partial \rho(x,t)}{\partial t} = \text{div}\left(\rho(x,t)\text{grad}f(x) + \text{grad}\rho(x,t)\right) = \text{div}\left(\rho(x,t)\text{grad}f(x)\right) + \Delta_M \rho(x,t), \quad (3)$$

where grad and div are gradient and divergence in Riemannian manifold, and Δ_M is the Laplace-Beltrami operator generalizing the Euclidean Laplacian Δ to Riemannian manifold. More details can be found in Appendix. The stationary solution of equation (3) is $e^{-f(x)}$ that minimizes the entropy regularized functional $\mathcal{L}(\rho)$, and then the optimization problem over the space of densities boils down to track the evolution of $\rho(x,t)$ that is defined by equation (3).

Coordinate-independent Langevin equation. In order to implement the aforementioned evolution of $\rho(x,t)$ in Euclidean space, one can simulate the stochastic process $\{X_t\}_{t\geq 0}$ defined by the Langevin equation: $dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t$, where B_t is the standard Brownian motion and X_t has $\rho(x,t)$ as its density function. In contrast to the Euclidean case, we need a coordinate-independent formulation of Langevin equation, e.g. Batrouni et al. [1986]. This is derived by expanding the Fokker-Planck equation (3) in a local coordinate system and is written in the following form:

$$dX_t = F(X_t)dt + \sqrt{2g^{-1}}dB_t \quad (4)$$

6

where $F(X_t)$ is a vector with i'th component $F_i = -\sum_j g^{ij} \frac{\partial f}{\partial x_j} + \frac{1}{\sqrt{|g|}} \sum_j \frac{\partial}{\partial x_j} \left(\sqrt{|g|} g^{ij} \right)$ and $|g|$ is the determinant of metric matrix g_{ij}. Note that this local form indicate the fact from Fokker-Planck equation that the process $\{X_t\}_{t \geq 0}$ is the negative gradient of f followed by a manifold Brownian motion. The rate of convergence we are interested in is the classic Euler-Maruyama discretization scheme in manifold setting, i.e. compute the vector in tangent space and project it onto the base manifold through exponential map. So the discretization error consists of two parts: one is from considering $\mathrm{grad} f(x_t)$ as constant in a neighborhood of x_t, and the other is from the approximation of a curved neighborhood of x_t with the tangent space at x_t. The main task in the proofs of Theorem 4.5 and 4.6 is to bound the aforementioned two parts of errors and compare with the density evolving along continuous time Langevin equation.

4.2 Convergence Analysis

We state some assumptions before presenting main theorems.

Assumption 4.1. M *is a closed manifold.*

It means M is compact and has no boundary, Definition 3.2. This assumption is essentially used to make the boundary integral on ∂M, i.e. $\int_{\partial M} \log \frac{\rho_t}{\nu} \langle \rho_t \mathrm{grad} f + \mathrm{grad} \rho_t, \mathbf{n} \rangle dx$ in the proof of Lemma 4.3, to vanish, see Appendix. This assumption can be relaxed to open manifold by assuming the integral decreases fast as x approaches the infinity.

Assumption 4.2. $f(x)$ *is differentiable on* M.

An immediate consequence by combining Assumption 4.1 and 4.2 is that there exists a number $L > 0$, such that the Riemannian gradient $\mathrm{grad} f$ of f is L-Lipschitz (Definition 3.8) due to the compactness of M. Another crucial property used in the proof is that the target distribution e^{-f} satisfies the log-Sobolev inequality, and this can also be derived by compactness of M.

Since the prerequisite of convergence of GLA is the convergence of the continuous time Langevin equation, we show the KL divergence between ρ_t and ν converges along the continuous time Riemannian Langevin equation. The proof is completed by the following lemma showing that $H(\rho_t|\nu)$ decreases since $\frac{d}{dt} H(\rho_t|\nu) < 0$ for all $\rho_t \neq \nu$. Based on the analysis of the previous section, it suffices to track the evolution of ρ_t according to the Fokker-Planck equation (3).

Lemma 4.3. *Suppose* ρ_t *evolves following the Fokker-Planck equation (3), then*

$$\frac{d}{dt} H(\rho_t|\nu) = -\int_M \rho_t(x) \left\| \mathrm{grad} \log \frac{\rho_t(x)}{\nu(x)} \right\|^2 dx$$

where dx *is the Riemannian volume element.*

The proof is a straightforward calculation of the time derivative of $H(\rho_t|\nu)$, followed by the expression of $\frac{\partial \rho_t}{\partial t}$ in equation (3), i.e.

$$\frac{d}{dt} H(\rho_t|\nu) = \int_M \frac{\partial \rho_t}{\partial t} \log \frac{\rho_t}{\nu} dx = \int_M \mathrm{div}(\rho_t \mathrm{grad} f + \mathrm{grad} \rho_t) \log \frac{\rho_t}{\nu} dx \tag{5}$$

$$= -\int_M \rho_t \left\| \mathrm{grad} \log \frac{\rho_t}{\nu} \right\|^2 dx + \int_{\partial M} \log \frac{\rho_t}{\nu} \langle \rho_t \mathrm{grad} f + \mathrm{grad} \rho_t, \mathbf{n} \rangle dx \tag{6}$$

The result follows from integration by parts and the Assumption 4.1. Details are left in Appendix.

Since M is compact, there exists a constant $\alpha > 0$ such that the log-Sobolev inequality (LSI) holds. So we can get the following convergence of KL divergence for continuous Langevin dynamics immediately.

Theorem 4.4. *Suppose* ν *satisfies LSI with constant* $\alpha > 0$. *Then along the Riemannian Langevin equation, i.e. the SDE (4) in local coordinate systems, the density* ρ_t *satisfies*

$$H(\rho_t|\nu) \leq e^{-2\alpha t} H(\rho_0|\nu).$$

The following theorem shows that the KL divergence $H(\rho_k|\nu)$ decreases geometrically along the GLA dynamics.

7

Theorem 4.5. *Suppose M is a closed manifold and R is the Riemann curvature (definition in Appendix), $\nu = e^{-f}$ a density on M with $\alpha > 0$ the log-Sobolev constant. Then there exists a global constant $K(R)$ determined by R, such that for any $x_0 \sim \rho_0$ with $H(\rho_0|\nu) \leq \infty$, the iterates $x_k \sim \rho_k$ of GLA with stepsize $\epsilon \leq \min\{\frac{\alpha}{4L\sqrt{L^2+K(R)}}, \frac{3L}{2K(R)}, \frac{1}{2L}, \frac{1}{2\alpha}\}$ satisfy*

$$H(\rho_k|\nu) \leq e^{-\frac{k}{2}\alpha\epsilon} H(\rho_0|\nu) + \frac{32\epsilon n L^2}{\alpha}$$

Proof Strategy: Recall for the Langevin algorithm in the Euclidean space Vempala and Wibisono [2019], the 1 step iteration (with initial point x_0) is given by $x_1 = x_0 - \epsilon \nabla f(x_0) + \sqrt{2\epsilon}\xi_0$, which can be interpreted as the numerical implementation of the solution of stochastic differential equation $dX_t = -\nabla f(x_0)dt + \sqrt{2}dB_t$ for time interval $t \in [0, \epsilon]$. The analogy for a gradient to be "constant" in a manifold is defined in that the gradient $\mathrm{grad} f(x)$ at x in a neighborhood of x_0 equals the parallel transport of $\mathrm{grad} f(x_0)$ to x, i.e. $\mathrm{grad} f(x) = \Gamma_{x_0}^x \mathrm{grad} f(x_0)$. Note that this defines a vector field on M except for a set of measure zero (with respect to the Riemannian volume measure), since there exists unique length minimizing geodesic between a given point x_0 and any point $x \in M$ except for the cut locus of x_0 that has measure of zero. The key step is to derive the partial differential eqation of the conditioned density of the stochastic differential equation with constant drift (constant with respect to the covariant derivative). The detailed proof is given in Appendix.

For the case where the curvature c is constant, the stepsize can be chosen specifically based on the curvature and Lipschitz constant of Riemannian gradient.

Theorem 4.6. *Suppose M is a closed manifold and c is the constant sectional curvature, $\nu = e^{-f}$ a density on M with $\alpha > 0$ the log-Sobolev constant. Then for stepsize $\epsilon \leq \min\{\frac{\sqrt{36}L^2}{|c|}, \frac{1}{2L}, \frac{1}{2\alpha}, \frac{\alpha}{\sqrt{32}L^2}\}$, for any $x_0 \sim \rho_0$ with $H(\rho_0|\nu) \leq \infty$, the iterates $x_k \sim \rho_k$ of GLA satisfy*

$$H(\rho_k|\nu) \leq e^{-\frac{k}{2}\alpha\epsilon} H(\rho_0|\nu) + \frac{64\epsilon n L^2}{3\alpha}.$$

The main idea of proving Theorem 4.6 follows that of Theorem 4.5, where the difference is using properties of Jacobi field (in constant curvature manifold) to obtain a better control for the choice of the stepsize. More information is provided in Appendix.

The convergence of KL divergence implies the convergence of Wasserstein distance. We denote $\frac{C\epsilon n L^2}{\alpha}$ the uniform error bound in Theorem 4.5 and 4.6 can the convergence of Wasserstein distance is stated as follows.

Proposition 4.7. *For the closed manifold M with a density $\nu = e^{-f(x)}$, the iterates $x_k \sim \rho_k$ of GLA with a properly chosen stepsize satisfy*

$$W_2(\rho_k, \nu)^2 \leq \frac{2}{\alpha} e^{-\frac{k}{2}\alpha\epsilon} H(\rho_0|\nu) + \frac{2C\epsilon n L^2}{\alpha^2}.$$

Proof. It is an immediate consequence from the convergence of KL divergence that the Wasserstein distance $W_2(\rho_k, \nu)$ converges rabidly since log-Sobolev inequality implies Talagrand inequality (Talagrand [1996], Otto and Villani [2000]). Since ν satisfies log-Sobolev inequality, then we have

$$W_2(\rho_k|\nu)^2 \leq \frac{2}{\alpha}H(\rho_k|\nu) \leq \frac{2}{\alpha}(e^{-\frac{k}{2}\alpha\epsilon} H(\rho_0|\nu) + \frac{C\epsilon n L^2}{\alpha}) = \frac{2}{\alpha} e^{-\frac{k}{2}\alpha\epsilon} H(\rho_0|\nu) + \frac{2C\epsilon n L^2}{\alpha^2} \quad (7)$$

and the proof completes. $\square$

5 Conclusion

In this paper we focus on the problem of sampling from a distribution on a Riemannian manifold and propose the Geodesic Langevin Algorithm. GLA modifies the Riemannian Langevin algorithm by using exponential map so that the algorithm is defined globally. By leveraging the geometric meaning of GLA, we provide a non-asymptotic convergence guarantee in the sense that the KL divergence (as well as the Wasserstein distance) decreases fast along the iterations of GLA. By assuming that we have full access to the geometric data (curvature, geodesic distance, Ricci curvature, diameter, etc) of the manifold, we can control the bias between the stationary distribution of GLA and the target distribution to be arbitrarily small through the choice of stepsize.

6 Broader Impact

Our results is motivated by variant applications of sampling algorithms in machine learning. Generating random sample especially in higher dimension with good convergence and error bound can be used in fast integration and volume computation. With the development of continuous game theory and GANs in recent years, our results have potential impact to solve the Nash equilibrium of the games with continuous strategy space.

References

Naman Agarwal, Nicolas Boumal, Brian Bullins, and Coralia Cartis. Adaptive regularization with cubics on manifolds. Mathematical Programming, 2020.

G.G. Batrouni, H. Kawai, and Pietro Rossi. Coordinate-independent formulation of the langevin equation. *Journal of Mathematical Physics*, 27, 1986.

Espen Bernton. Langevin monte carlo and jko splitting. In *COLT*, 2018.

Simon Byrne and Mark Girolami. Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics, Theory and Applications*, 40(4), 2013.

NS. Chatterji, J. Diakonikolas, MI. Jordan, and Peter Bartlett. Langevin monte carlo without smoothness. In *arXiv:1905.13285*, 2019.

Xiang Cheng and Peter Bartlett. Convergence of langevin mcmc in kl-divergence. *Proceedings of Machine Learning Research*, 83, 2018.

Xiang Cheng, Niladri S Chatterji, Yasin Abbasi-Yadkori, Peter Bartlett, and Michael I Jordan. Sharp convergence rates for langevin dynamics in the nonconvex setting. *arXiv:1805.01648*, 2018.

Arnak Dalalyan. Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. *Proceedings of the Conference on Learning Theory*, 2017a.

Arnak Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3), 2017b.

Arnak Dalalyan and Avetik Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 2019.

Carles Domingo-Enrich, Samy Jelassi, Arthur Mensch, Grand Rotskoff, and Joan Bruna. A mean-field analysis of two-player zero-sum games. https://arxiv.org/abs/2002.06277, 2020.

Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted langevin algorithm. *The Annals of Applied Probability*, 27(3), 2017.

Alain Durmus and Eric Mounline. High-dimensional bayesian inference via the unadjusted langevin algorithm. *Bernoulli*, 25(4A), 2019.

Alain Durmus, Eric Moulines, and Eero Saksman. On the convergence of hamiltonian monte carlo. In *arXiv:1705.00166*, 2017.

Alain Durmus, Szymon Majewski, and Blazej Miasojedow. Analysis of langevin monte carlo via convex optimization. In *arXiv:1802.09188*, 2018.

Raaz Dwivedi, Yuansi Chen, Martin Wainwright, and Bin Yu. Log-concave sampling: Metropolis-hastings algorithms are fast! *Proceedings of the Conference of Learning Theory*, 2018.

Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Coupling and quantitative contraction rates for langevin dynamics. In *arXiv:1703.01617*, 2018.

Rong Ge, Holden Lee, and Andrej Risteski. Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering langevin monte carlo. In *NeurIPS*, 2018.

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), 2011.

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *ICML*, 2017.

Leonard Gross. Logarithmic sobolev inequalities and contractivity properties of semigroups. *Lecture Notes in Maths*, 1563, 1993.

Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. Mirrored langevin dynamics. In *NeurIPS*, 2018.

Elton P. Hsu. *Stochastic Analysis on Manifolds*. American Mathematical Society, 2002.

Alfredo Garbuno Inigo, Nikolas Nüsken, and Sebastian Reich. Affine invariant interacting langevin dynamics for bayesian. In *arXiv:1912.02859*, 2019.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker-planck equation. *SIAM Journal on Mathematical Analysis*, 29(1), 1998.

R. Kannan, L. Lovász, and M. Simonovits. Random walks and an $o^*(n^5)$ volume algorithm for convex bodies. *Random Structures and Algorithms*, 11, 1997.

Michel Ledoux. Concentration of measure and logarithmic sobolev inequalities. *Séminaire de probabilités*, 33, 1999.

John Lee. *Introduction to Riemannian Manifolds*, volume 176 GTM. Springer, 2018.

Chang Liu, Jun Zhu, and Yang Song. Stochastic gradient geodesic mcmc methods. In *NIPS*, 2016.

L. Lovász and S. Vempala. Fast algorithm for logconcave functions: sampling, rounding, integration and optimization. In *FOCS*, 2006a.

L. Lovász and S. Vempala. Hit-and-run from a corner. *SIAM Journal on Computing*, 35(4), 2006b.

L. Lovász and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures and Algorithms*, 30(3), 2007.

Tung Duy Luu, Jalal Fadili, and Christophe Chesneau. Sampling from non-smooth distribution through langevin diffution. *URL https://hal.archives-ouvertes.fr/hal-01492056*, 2017.

Yi-An Ma, Niladri Chatterji, Xiang Cheng, Nicolas Flammarion, Peter Bartlett, and Michael I Jordan. Is there an analog of nesterov acceleration for mcmc? In *arXiv preprint arXiv: 1902.00996*, 2019.

Mateusz Majka, Aleksandar Mijatović, and Lukasz Szpruch. Non-asymptotic bounds for sampling algorithms without logconcavity. *arXiv:1808.07105*, 2018.

Oren Mangoubi and Nisheeth Vishnoi. Dimensionally tight bounds for second-order hamiltonian monte carlo. In *NeurIPS*, 2018.

Ankur Moitra and Andrej Risteski. Fast convergence for langevin diffusion with matrix manifold structure. In *arXiv:2002.05576*, 2020.

Felix Otto and Cedric Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173:361–400, 2000.

Sam Patterson and Yee Whye Teh. Stochastic gradient riemannian langevin dynamics on the probability simplex. In *NIPS*, 2013.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *COLT*, 2017.

Lillian Ratliff, Samuel Burden, and S. Shankar Sastry. Characterization and computation of local nash equilibria in continuous games. In *Fifty-first Annual Allerton Conference*, 2013.

Lillian Ratliff, Samuel Burden, and S. Shankar Sastry. On the characterization of local nash equilibria in continuous games. *IEEE Transactions on Automatic Control*, 61(8), 2016.

Gareth Roberts and Osnat Stramer. Langevin diffusions and metropolis-hastings algorithms. *Methodology and computing in applied probability*, 4(4), 2002.

Gareth Roberts and Richard Tweedie. Exponential convergence of langevin distributions and their discrete approximation. *Bernoulli*, 2(4), 1996.

O Rothaus. Diffusion on compact riemannian manifolds and logarithmic sobolev inequalities. *Journal of Functional Analysis*, 42, 1981.

O Rothaus. Hypercontractivity and the bakry-emery criterion. *Journal of Functional Analysis*, 65, 1986.

Filippo Santanmbrogio. *Optimal Transport for Applied Mathematicians*. Birkhäuser, 2015.

Christof Seiler, Simon Rubinstein-Salzedo, and Susan Holmes. Positive curvature and hamiltonian monte carlo. In *NIPS*, 2014.

Samuel L. Smith, Daniel Duckworth, Semon Rezchikov, Quoc V. Le, and Jascha Sohl-Dickstein. Stochastic natural gradient descent draws posterior samples in function space. In *NeurIPS*, 2018.

M Talagrand. Transportation cost for gaussian and other product measures. *Geometric and Functional Analysis*, 6, 1996.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In *NeurIPS*, 2019.

Santosh S. Vempala and Yin-Tat Lee. Geodesic walks in polytopes. In *STOC*, 2017.

Santosh S. Vempala and Yin-Tat Lee. Convergence rate of riemannian hamiltonian monte carlo and faster polytope volume computation. In *STOC*, 2018.

Feng-Yu Wang. On estimation of the logarithmic sobolev constant and gradient estimates of heat semigroups. *Probability Theory and Related Fields*, 108:87–101, 1997a.

Feng-Yu Wang. Logarithmic sobolev inequalities on noncompact riemannian manifolds. *Probability Theory and Related Fields*, 109:417–424, 1997b.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *ICML*, 2011.

Andre Wibisono. Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, 2018.

Andre Wibisono. Proximal langevin algorithm: Rapid convergence under isoperimetry. In *arXiv:1911.01469*, 2019.

Kelvin Shuangjian Zhang, Gabriel Peyré, Jalal Fadili, and Marcelo Pereyra. Wasserstein control of mirror langevin monte carlo. *arXiv:2002.04363*, 2020.

A   More background

A.1   Calculus on manifold

Definition A.1 (Levi-Civita Connection). Let (M, g) be a Riemannian manifold. An affine connection is said to be the Levi-Civita connection if it is torsion-free. i.e.

$$\nabla_X Y - \nabla_Y X = [X, Y]$$

for every pair of vector fields X, Y on M and preserves the metric i.e.

$$\nabla g = 0.$$

Definition A.2 (Riemannian Curvature Tensor). Let (M, g) be a Riemannian manifold with the associated Levi-Civita connection ∇. Define the Riemannian curvature tensor as

$$R(u, v)w = \nabla_u \nabla_v w - \nabla_v \nabla_u w - \nabla_{[u,v]} w.$$

where u, v are vector fields on the M.

Definition A.3 (Sectional Curvature). Let (M, g) be a Riemannian manifold with the Riemannian curvature tensor R. The sectional curvature K is

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

where u, v are linearly independent vectors on the tangent space at a point on the manifold.

Definition A.4 (Ricci Curvature). Let (M, g) be a Riemannian manifold with the Riemannian curvature tensor R. The Ricci curvature tensor Ric_M is the trace of the linear map

$$X \rightarrow R(X, Y)Z.$$

where X, Y, Z are vector fields on M.

Note the Ricci curvature is not used in the proof but used in presenting the estimate of the log-Sobolev constant.

The following results on Jacobi field is used in proving Theorem 4.6.

Definition A.5 (Variation through geodesic). Let $I, K \subset \mathbb{R}$ are intervals, $\gamma : I \rightarrow M$ is a geodesic, and $\Gamma : K \times I \rightarrow M$ is a variation of γ. We say that Γ is a variation through geodesic if each of the main curves $\Gamma_s(t) = \Gamma(s, t)$ is also a geodesic.

Theorem A.6 (The Jacobi Equation). *Let (M, g) be a Riemannian manifold, let γ be a geodesic, and let $J(t)$ be a vector field along γ. If J is the variation field of a variation through geodesic, then $J(t)$ satisfies the following equation, called the Jacobi equation:*

$$D_t^2 J(t) + R(J(t), \gamma'(t))\gamma'(t) = 0.$$

Definition A.7 (Jacobi Field). We call a smooth vector field along a geodesic that satisfies the Jacobi equation a Jacobi field.

Definition A.8 (Riemannian Volume). Let (M, g) be an orientable Riemannian manifold. The volume form on the manifold in local coordinates is given as

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

We denote $|g| = \det(g)$ and $dx = d\text{Vol}$ (if no ambiguities caused) for short throughout following context. The following Theorem is used to guarantee the exponential map is defined on the whole tangent space, which is equivalent to require M to be complete. This property is satisfied in our setting for M to be compact without boundary.

Theorem A.9 (Hopf-Rinow). *Let (M, g) be a connected Riemannian manifold. Then the followings are equivalent.*

1. *The closed and bounded subsets of M are compact.*

2. *M is a complete metric space.*

3. *M is geodesically complete: for every point $x \in M$, the exponential map Exp_x is defined on the entire tangent space $T_x M$.*

The notion of differential operators, e.g. gradient, divergence and Laplacian for the differentiable functions and vector fields on Euclidean space can be generalized to Riemannian manifold. In local coordinate system, $\{\partial_i = \frac{\partial}{\partial x_i} : i \in [n]\}$ is a basis of the tangent space $T_x M$. Denote g_{ij} the metric matrix, g^{ij} the inverse of g_{ij}

and $|g| = \det g_{ij}$ the determinant of matrix g_{ij}. Let f and V be differentiable function and vector field on M, then the Riemannian gradient of f and the divergence of V are written as

$$\mathrm{grad} f = \sum_{i,j} g^{ij} \frac{\partial f}{\partial x_i} \partial_i \quad \text{and} \quad \mathrm{div} V = \frac{1}{\sqrt{|g|}} \sum_i \frac{\partial}{\partial x_i} \left(\sqrt{|g|} V_i \right)$$

where V_i is the i-th component of V.

The Laplace-Beltrami operator Δ_M acting on f is defined to be the divergence of the gradient of f, i.e.

$$\Delta_M f = \mathrm{div}(\mathrm{grad} f) = \frac{1}{\sqrt{|g|}} \sum_i \frac{\partial}{\partial x_i} \left(\sqrt{|g|} \sum_j g^{ij} \frac{\partial f}{\partial x_j} \right).$$

In Euclidean space, Δ_M boils down to the classic Laplacian $\Delta f = \nabla \cdot (\nabla f)$.

The following integration by parts formulas are used in proof of main lemmas. Let M be a compact oriented Riemannian manifold of dimension n with boundary ∂M. Let X be a vector field on M. The integration by parts is given by

$$\int_M \langle \mathrm{grad} f, X \rangle = -\int_M f \mathrm{div} X + \int_{\partial M} f \langle X, n \rangle$$

or Green's formula

$$\int_M (f \Delta_M g - g \Delta_M f) = -\int_{\partial M} \left(f \frac{\partial g}{\partial n} - g \frac{\partial f}{\partial n} \right)$$

If ∂M is empty or the vector field X decay sufficiently fast at infinity of M provided M is open, we have

$$\int_M \langle \mathrm{grad} f, X \rangle = -\int_M f \mathrm{div} X.$$

Definition A.10 (First eigenvalue of Laplacian). The first eigenvalue $\lambda_1 \geq 0$ of the Laplacian operator on M is defined to be

$$\lambda_1 = \inf_{f \in C_c^\infty} \left\{ \frac{\int_M \|\mathrm{grad} f\|^2 dx}{\int_M \|f\|^2 dx} \right\}.$$

A.2 Stochastic analysis on manifold

Recall that the standard Brownian motion in $\mathbb{R}^n$ is a random process $\{X_t\}_{t \geq 0}$ whose density evolves according to the diffusion equation

$$\frac{\partial \rho(x,t)}{\partial t} = \frac{1}{2} \Delta \rho(x,t).$$

Similarly, the Brownian motion in manifold M is M-valued random process $\{W_t\}_{t \geq 0}$ whose density function evolves according to the diffusion equation with respect to Laplace-Beltrami operator which is the counterpart of the Laplace operator on Euclidean space.

$$\frac{\partial \rho(x,t)}{\partial t} = \frac{1}{2} \Delta_M \rho(x,t).$$

In local coordinate, the Laplace-Beltrami is written as

$$\Delta_M = \sum_{i,j} g^{ij} \frac{\partial^2}{\partial x_i \partial x_j} + \sum_i b_i \frac{\partial}{\partial x_i},$$

where

$$b_i = \sum_j \frac{1}{\sqrt{|g|}} \frac{\partial}{\partial x_j} \left(\sqrt{|g|} g^{ij} \right) = \sum_{j,k} g^{jk} \Gamma_{jk}^i. \tag{8}$$

We can construct Brownian motion in the local coordinate as the solution of the stochastic differential equation for a process $\{X_t\}_{t \geq 0}$:

$$dX_t = \frac{1}{2} b(X_t) dt + \sigma(X_t) dB_t$$

where the component $b_i(X_t)$ of $b(X_t)$ is given by (8) and $\sigma = (\sigma_{ij})$ is the unique symmetric square root of $g^{-1} = (g^{ij})$.

B Derivation of the GLA

In this section, we give detailed explanation on that the Riemannian Langevin algorithm, as a stochastic process, captures the dynamics of the evolution of the density function for the stochastic process. The derivation is firstly

13

to write the diffusion equation in local coordinate system of the manifold, and then compare the corresponding terms to the Fokker-Planck equation related to stochastic differential equation that gives insight to the local expression of Riemannian Langevin algorithm. In order to do this, recall that the density e^{-f} on M is the stationary solution of the PDE

$$\frac{\partial \rho_t}{\partial t} = \text{div}\left(\rho_t \text{grad} f + \text{grad} \rho_t\right). \tag{9}$$

Using the local expression of Riemannian gradient and divergence operator, this PDE can be written as

$$\frac{\partial \rho_t}{\partial t} = \frac{1}{\sqrt{|g|}} \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \left(\sqrt{|g|}\left(\sum_j g^{ij} \frac{\partial f}{\partial x_j}\right)\rho_t + \sqrt{|g|}\sum_j g^{ij}\frac{\partial \rho_t}{\partial x_j}\right) \tag{10}$$

$$= \frac{1}{\sqrt{|g|}} \sum_i \frac{\partial}{\partial x_i}\left(\left(\sum_j g^{ij}\frac{\partial f}{\partial x_j} - \frac{1}{\sqrt{|g|}}\sum_j \frac{\partial}{\partial x_j}\left(\sqrt{|g|}g^{ij}\right)\right)\sqrt{|g|}\rho_t\right) \tag{11}$$

$$+ \frac{1}{\sqrt{|g|}}\sum_{i,j}\frac{\partial^2}{\partial x_i \partial x_j}\left(g^{ij}\sqrt{|g|}\rho_t\right) \tag{12}$$

Denoting $\tilde{\rho}_t = \sqrt{|g|}\rho_t$, we have the Fokker-Planck equation of density in Euclidean space as follows,

$$\frac{\partial \tilde{\rho}_t}{\partial t} = -\sum_i \frac{\partial}{\partial x_i}\left(\left(\frac{1}{\sqrt{|g|}}\sum_j \frac{\partial}{\partial x_j}\left(\sqrt{|g|}g^{ij}\right) - \sum_j g^{ij}\frac{\partial f}{\partial x_j}\right)\tilde{\rho}_t\right) + \sum_{i,j}\frac{\partial^2}{\partial x_i \partial x_j}\left(g^{ij}\tilde{\rho}_t\right). \tag{13}$$

Since for any stochastic differential equation of the form

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

the density p_t for X_t satisfies

$$\frac{\partial p(x,t)}{\partial t} = -\sum_{i=1}^{n}\frac{\partial}{\partial x_i}\left(F_i(x,t)p(x,t)\right) + \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial^2}{\partial x_i \partial x_j}\left(A_{ij}(x,t)p(x,t)\right) \tag{14}$$

where $A = \frac{1}{2}\sigma\sigma^\top$, i.e. $A_{ij} = \frac{1}{2}\sum_{k=1}^{n}\sigma_{ik}(x,t)\sigma_{jk}(x,t)$. Compare equations (13) and (14), we have the drift and diffusion terms in local coordinate systems are given by

$$F_i(x_t) = -\sum_j g^{ij}\frac{\partial f}{\partial x_j} + \frac{1}{\sqrt{|g|}}\sum_j \frac{\partial}{\partial x_j}\left(\sqrt{|g|}g^{ij}\right)$$

and

$$\sigma(x_t) = \sqrt{2(A_{ij})} = \sqrt{2(g^{ij})} = \sqrt{2g^{-1}}.$$

So the local Langevin equation is

$$dX_t = F(X_t)dt + \sqrt{2g^{-1}}dB_t. \tag{15}$$

This equation describes infinitesimal evolution of X_t, which can be seen as a process in the tangent space of M. The Riemannian Langevin algorithm is the classic Euler-Maruyama discretization in the tangent space, i.e., by letting X_t move in the tangent space for a positive time interval $t \in [0, \epsilon]$ with the drift and diffusion at current location. Suppose the initial point is x_0, the tangent vector is

$$\epsilon F(x_0) + \sqrt{2\epsilon g^{-1}(x_0)}\xi_0$$

where $\xi_0 \sim \mathcal{N}(0, I)$ is the standard Gaussian noise. Then the updated point is obtained by mapping the vector to the base manifold via exponential map,

$$x_1 = \text{Exp}_{x_0}\left(\epsilon F(x_0) + \sqrt{2\epsilon g^{-1}(x_0)}\xi_0\right).$$

Renaming $x_k = x_0$ and $x_{k+1} = x_1$, we have the general form

$$x_{k+1} = \text{Exp}_{x_k}\left(\epsilon F(x_k) + \sqrt{2\epsilon g^{-1}(x_k)}\xi_0\right).$$

We give the expression of the algorithm in normal coordinate, for convenience in part of the proofs of main theorems.

For any manifold M, and $x \in M$, $T_x M$ is isomorphic to $\mathbb{R}^n$, $\exp_x^{-1}$ gives a local coordinate system of M around x. This is called the normal coordinates at x. The following lemmas are from Lee-Vampalar

Lemma B.1. *In normal coordinate, we have*

$$g_{ij}(x) = \delta_{ij} - \frac{1}{3}\sum_{kl}R_{ikjl}(x)x^k x^l + O(|x|^3).$$

Under normal coordinate, the RLA can be written as

$$x_{t+1} = \text{Exp}_{x_t}(-\epsilon\nabla f(x_t) + \sqrt{2\epsilon}\xi_0).$$

Note that the expression in the tangent space is exactly the same as unadjusted Langevin algorithm in Euclidean space.

14

C Missing proofs of Section 4

C.1 Proof of Theorem 4.4

In this section, we proof that the KL divergence decreases along the process evolving following Riemannian Langevin equation.

Firstly, need show that according to the SDE on manifold in local chart, the density function evolves according to Fokker-Planck/diffusion equation on this manifold.

Lemma 4.3. Suppose ρ_t evolves following the Fokker-Planck equation (3), then

$$\frac{d}{dt} H(\rho_t|\nu) = -\int_M \rho_t(x) \left\| \operatorname{grad} \log \frac{\rho_t(x)}{\nu(x)} \right\|^2 dx$$

where dx is the Riemannian volume element.

Proof. Since

$$\int_M \rho_t \frac{\partial}{\partial t} \log \frac{\rho_t}{\nu} dx = \int_M \rho_t \frac{\partial}{\partial t} (\log \rho_t + f) dx \tag{16}$$

$$= \int_M \frac{\partial \rho_t}{\partial t} dx \tag{17}$$

$$= \frac{d}{dt} \int_M \rho_t dx = 0, \tag{18}$$

we have

$$\frac{d}{dt} H(\rho_t|\nu) = \frac{d}{dt} \int_M \rho_t \log \frac{\rho_t}{\nu} dx \tag{19}$$

$$= \int_M \frac{d}{dt} \left(\rho_t \log \frac{\rho_t}{\nu} \right) dx \tag{20}$$

$$= \int_M \frac{\partial \rho_t}{\partial t} \log \frac{\rho_t}{\nu} dx + \int_M \rho_t \frac{\partial}{\partial t} \log \frac{\rho_t}{\nu} dx \tag{21}$$

$$= \int_M \frac{\partial \rho_t}{\partial t} \log \frac{\rho_t}{\nu} dx. \tag{22}$$

Plug in with diffusion equation

$$\frac{\partial \rho_t}{\partial t} = \operatorname{div}(\rho_t \operatorname{grad} f + \operatorname{grad} \rho_t)$$

and apply integration by parts, we obtain

$$\frac{d}{dt} H(\rho_t|\nu) = \int_M \operatorname{div}(\rho_t \operatorname{grad} f + \operatorname{grad} \rho_t) \log \frac{\rho_t}{\nu} dx \tag{23}$$

$$= -\int_M \rho_t \left\| \operatorname{grad} \log \frac{\rho_t}{\nu} \right\|^2 dx \tag{24}$$

$$+ \int_{\partial M} \log \frac{\rho_t}{\nu} \langle \rho_t \operatorname{grad} f + \operatorname{grad} \rho_t, n \rangle dx \tag{25}$$

Since M is compact and has no boundary, the boundary integral equals to zero, then we have

$$\frac{d}{dt} H(\rho_t|\nu) = -\int_M \rho_t \left\| \operatorname{grad} \log \frac{\rho_t}{\nu} \right\|^2 dx$$

$\square$

Theorem 4.4. Suppose ν satisfies LSI with constant $\alpha > 0$. Then along the Riemannian Langevin equation, i.e. the SDE (4) in local coordinate systems, the density ρ_t satisfies

$$H(\rho_t|\nu) \le e^{-2\alpha t} H(\rho_0|\nu).$$

Proof. By LSI, we have

$$\frac{d}{dt} H(\rho_t|\nu) \le -2\alpha H(\rho_t|\nu),$$

multiplying both sides by $e^{\alpha t}$,

$$e^{2\alpha t} \frac{d}{dt} H(\rho_t|\nu) \le -2\alpha e^{2\alpha t} H(\rho_t|\nu)$$

15

and then

$$e^{2\alpha t}\frac{d}{dt}H(\rho_t|\nu) + 2\alpha e^{2\alpha t}H(\rho_t|\nu) = \frac{d}{dt}e^{2\alpha t}H(\rho_t|\nu) \le 0.$$

Integrating for $0 \le t \le s$, the result holds as

$$e^{2\alpha s}H(\rho_s|\nu) - H(\rho_0|\nu) = \int_0^s \frac{d}{dt}e^{2\alpha t}H(\rho_t|\nu) \le 0.$$

Rearranging and renaming s by t, we conclude

$$H(\rho_t|\nu) \le e^{-2\alpha t}H(\rho_t|\nu)$$

$\square$

C.2 Proof of Theorem 4.5

Lemma C.1. *Assume $\nu = e^{-f}$ is L-smooth. Then*

$$\mathbb{E}_\nu[\|\mathrm{grad}f\|^2] \le nL.$$

Proof. Since

$$\mathbb{E}_\nu[\|\mathrm{grad}f\|^2] = \int_M \langle \mathrm{grad}f, \mathrm{grad}f\rangle e^{-f}dx = -\int_M \langle \mathrm{grad}e^{-f}, \mathrm{grad}f\rangle dx,$$

where dx is the Riemannian volume element. Integration by parts on manifold gives the following

$$-\int_M \langle \mathrm{grad}e^{-f}, \mathrm{grad}f\rangle dx = \int_M e^{-f}\Delta_M f dx - \int_{\partial M} e^{-f}\langle \mathrm{grad}f, n\rangle ds$$

where ds is the area element on ∂M. By the assumption that M is boundaryless, the integral on the boundary is 0. By the assumption $\mathrm{Hess}f$ is L-smooth and the fact that $\mathrm{Hess}f \ge \frac{1}{n}\Delta_M f$, we conclude $\mathbb{E}_\nu[\|\mathrm{grad}f\|^2] \le nL$ $\square$

Lemma C.2. *Suppose ν satisfies Talagrand inequality with constant $\alpha > 0$ and L-smooth. Then for any ρ,*

$$\mathbb{E}_\rho[\|\mathrm{grad}f\|^2] \le \frac{4L^2}{\alpha}H(\rho|\nu) + 2nL.$$

Proof. Let $x \sim \rho$ and $y \sim \nu$ with optimal coupling (x, y) so that

$$\mathbb{E}[d(x,y)^2] = W_2(\rho,\nu)^2.$$

$\mathrm{grad}f$ is L-Lipschitz from the assumption that f is L-smooth. So we have the following inequality:

$$\|\mathrm{grad}f(x)\| \le \|\mathrm{grad}f(x) - \Gamma_y^x\mathrm{grad}f(y)\| + \|\Gamma_y^x\mathrm{grad}f(y)\| \tag{26}$$
$$\le Ld(x,y) + \|\Gamma_y^x\mathrm{grad}f(y)\| \tag{27}$$
$$= Ld(x,y) + \|\mathrm{grad}f(y)\| \tag{28}$$

where the equality follows from that parallel transport is an isometry. The same arguments as V-W gives

$$\|\mathrm{grad}f(x)\|^2 \le (Ld(x,y) + \|\mathrm{grad}f(y)\|)^2 \le 2Ld(x,y)^2 + 2\|\mathrm{grad}f(y)\|^2$$

and

$$\mathbb{E}_\rho[\|\mathrm{grad}f(x)\|^2] \le 2L^2\mathbb{E}[d(x,y)^2] + 2\mathbb{E}_\nu[\|\mathrm{grad}f(y)\|^2] \tag{29}$$
$$= 2L^2W_2(\rho,\nu)^2 + 2\mathbb{E}_\nu[\|\mathrm{grad}f(y)\|^2]. \tag{30}$$

By Talagrand inequality and previous lemma, the result follows. $\square$

Lemma C.3. *Suppose ν satisfies LSI with constant $\alpha > 0$ and is L-smooth. If ϵ small enough, then along each step,*

$$H(\rho_\epsilon|\nu) \le e^{-\frac{1}{2}\alpha\epsilon}H(\rho_0|\nu) + 6\epsilon^2nL^2 + 4K(R)nL\epsilon^3$$

for small ϵ, and

$$H(\rho_{k+1}|\nu) \le e^{-\frac{1}{2}\alpha\epsilon}H(\rho_k|\nu) + 6\epsilon^2nL^2 + 4K(R)nL\epsilon^3$$

for all $k \in \mathbb{N}$, where $K(R)$ is global constant depending on the Riemann curvature R.

Proof. Suppose $x_0 \sim \rho_0$, we consider the local Langevin equation derived in previous section. The Riemannian Langevin equation can be written in terms of gradient $(\text{grad} f)$ and Brownian motion (W_t) on Riemannian manifold, in normal coordinate system, it has the form of

$$dx_t = -\nabla f(x_t)dt + \sqrt{2}dB_t.$$

For the algorithm, we consider $\text{grad} f(x_t)$ is constant in neighborhood of x_0. By constant we mean the covariant derivative $\nabla_z \text{grad} f(x_t) = 0$ for any directional vector $z \in T_{x_0}M$, i.e. $\text{grad} f(x_t) = \text{grad} f(x_0)$ when $t = 0$, and $\text{grad} f(x_t) = \Gamma_{x_0}^{x_t} \text{grad} f(x_0)$. We next derive the continuity equation for the SDE

$$dx_t = -\nabla f(x_0)dt + \sqrt{2}dB_t.$$

In this case, we actually assume the vector field in tangent space is constant, and the constant vector field $\text{grad} f(x_0)$ induces a vector field $V(x)$ on a neighborhood of x_0 on manifold by exponential map,

$$V(x_0, x) = D\text{Exp}_{x_0}(\text{Exp}_{x_0}^{-1}(x))[\text{grad} f(x_0)].$$

where $D\text{Exp}_{x_0}(v)[\cdot]$ denotes the Jacobian of the exponential map at $v \in T_{x_0}M$. Since the domain of exponential map is $T_{x_0}M$ that is Euclidean, $\text{grad} f(x_0) \in T_{x_0}M$ is also in the tangent space at $\text{Exp}_{x_0}^{-1}(x)$. For each $t > 0$, let $\rho_{0t}(x_0, x_t)$ denote the joint distribution of (x_0, x_t), and we have that

$$\rho_{0t}(x_0, x_t) = \rho_0(x_0)\rho_{t|0}(x_t, x_0) = \rho_t(x_t)\rho_{0|t}(x_0, x_t).$$

The diffusion equation conditioning on x_0 is the following

$$\frac{\partial \rho_{t|0}(x_t|x_0)}{\partial t} = \text{div}\left(\rho_{t|0}(x_t|x_0)V(x_t) + \text{grad}\rho_{t|0}(x_t|x_0)\right) \tag{31}$$

$$= \text{div}\left(\rho_{t|0}(x_t|x_0)V(x_t)\right) + \Delta_M \rho_{t|0}(x_t|x_0). \tag{32}$$

Multiplying both sides by $\rho_0(x_0)$, we have

$$\rho_0(x_0)\frac{\partial \rho_{t|0}(x_t|x_0)}{\partial t} = \left(\text{div}\left(\rho_{t|0}(x_t|x_0)V(x_t)\right) + \Delta_M \rho_{t|0}(x_t|x_0)\right)\rho_0(x_0).$$

Since $\rho_t(x)$ is the marginal distribution of $\rho_{0t}(x_0, x)$, so

$$\rho_t(x) = \int_M \rho_{0t}(x_0, x)dx_0 \tag{33}$$

$$= \int_M \rho_{t|0}(x|x_0)\rho_0(x_0)dx_0. \tag{34}$$

Then

$$\frac{\partial \rho_t(x)}{\partial t} = \int_M \frac{\partial \rho_{t|0}(x|x_0)}{\partial t}\rho_0(x_0)dx_0 \tag{35}$$

$$= \int_M \left(\text{div}\left(\rho_{t|0}(x|x_0)V(x_0, x)\right) + \Delta_M \rho_{t|0}(x|x_0)\right)\rho_0(x_0)dx_0 \tag{36}$$

$$= \int_M \left(\text{div}\left(\rho_{t|0}(x|x_0)\rho_0(x_0)V(x_0, x)\right) + \Delta_M \rho_{t|0}(x|x_0)\rho_0(x_0)\right)dx_0 \tag{37}$$

$$= \int_M \text{div}\left(\rho_t(x)\rho_{0|t}(x_0|x)V(x_0, x)\right)dx_0 + \int_M \Delta_M \rho_{t0}(x_0, x)dx_0 \tag{38}$$

$$= \text{div}\left(\rho_t(x)\int_M \rho_{0|t}(x_0|x)V(x_0, x)dx_0\right) + \Delta_M \rho_t(x) \tag{39}$$

$$= \text{div}\left(\rho_t(x)\mathbb{E}_{\rho_{0|t}}[V(x_0, x)|x_t = x]\right) + \Delta_M \rho_t(x). \tag{40}$$

Since $\int_M \frac{\partial \rho_t}{\partial t}dx = 0$, we have

$$\frac{d}{dt}H(\rho_t|\nu) = \int_M \frac{\partial \rho_t}{\partial t}\log\frac{\rho_t}{\nu}dx.$$

17

Substituting,

$$\frac{d}{dt} H(\rho_t|\nu) = \int_M \left(\text{div} \left(\rho_t(x) \mathbb{E}_{\rho_{0|t}}[V(x_0, x)|x_t = x] \right) + \Delta_M \rho_t(x) \right) \log \frac{\rho_t}{\nu} dx \tag{41}$$

$$= \int_M \left[\text{div} \left(\rho_t(x) \mathbb{E}_{\rho_{0|t}}[V(x_0, x)|x_t = x] + \text{grad} \rho_t(x) \right) \right] \log \frac{\rho_t}{\nu} dx \tag{42}$$

$$= \int_M \left[\text{div} \left(\rho_t(x) \mathbb{E}_{\rho_{0|t}}[V(x_0, x)|x_t = x] + \rho_t(x) \frac{\text{grad} \rho_t(x)}{\rho_t(x)} \right) \right] \log \frac{\rho_t}{\nu} dx \tag{43}$$

$$= \int_M \left[\text{div} \left(\rho_t(x) \left(\mathbb{E}_{\rho_{0|t}}[V(x_0, x)|x_t = x] + \text{grad} \log \rho_t(x) \right) \right) \right] \log \frac{\rho_t}{\nu} dx \tag{44}$$

$$= \int_M \left[\text{div} \left(\rho_t(x) \left(\mathbb{E}_{\rho_{0|t}}[V(x_0, x)|x_t = x] + \text{grad} \log \rho_t(x) - \text{grad} \log e^{-f} - \text{grad} f(x) \right) \right) \right] \log \frac{\rho_t}{\nu} dx \tag{45}$$

$$= \int_M \left[\text{div} \left(\rho_t(x) \left(\mathbb{E}_{\rho_{0|t}}[V(x_0, x)|x_t = x] + \text{grad} \log \frac{\rho_t}{\nu} - \text{grad} f(x) \right) \right) \right] \log \frac{\rho_t}{\nu} dx \tag{46}$$

$$= -\int_M \rho_t \left\langle \left(\text{grad} \log \frac{\rho_t}{\nu} + \mathbb{E}_{\rho_{0|t}}[V(x_0, x)|x_t = x] - \text{grad} f(x) \right), \text{grad} \log \frac{\rho_t}{\nu} \right\rangle dx \tag{47}$$

$$= -\int_M \rho_t \left\| \text{grad} \log \frac{\rho_t}{\nu} \right\|^2 dx + \int_M \rho_t \left\langle \text{grad} f(x) - \mathbb{E}_{\rho_{0|t}}[V(x_0, x)|x_t = x], \text{grad} \log \frac{\rho_t}{\nu} \right\rangle dx \tag{48}$$

where

$$\left\langle \text{grad} f(x) - \mathbb{E}_{\rho_{0|t}}[V(x_0, x)|x_t = x], \text{grad} \log \frac{\rho_t}{\nu} \right\rangle \tag{49}$$

$$= \left\langle \text{grad} f(x) - \int_M \rho_{0|t}(x_0|x) V(x_0, x) dx_0, \text{grad} \log \frac{\rho_t}{\nu} \right\rangle \tag{50}$$

$$= \left\langle \int_M \text{grad} f(x) \rho_{0|t}(x_0|x) dx_0 - \int_M \rho_{0|t}(x_0|x) V(x_0, x) dx_0, \text{grad} \log \frac{\rho_t}{\nu} \right\rangle \tag{51}$$

$$= \left\langle \int_M \left(\text{grad} f(x) \rho_{0|t}(x_0|x) - \rho_{0|t}(x_0|x) V(x_0, x) \right) dx_0, \text{grad} \log \frac{\rho_t}{\nu} \right\rangle \tag{52}$$

$$= \left\langle \int_M \left(\text{grad} f(x) - V(x_0, x) \right) \rho_{0|t}(x_0|x) dx_0, \text{grad} \log \frac{\rho_t}{\nu} \right\rangle \tag{53}$$

$$= \int_M \left\langle \text{grad} f(x) - V(x_0, x), \text{grad} \log \frac{\rho_t}{\nu} \right\rangle \rho_{0|t}(x_0|x) dx_0. \tag{54}$$

So

$$\int_M \rho_t \left\langle \text{grad} f(x) - \mathbb{E}_{\rho_{0|t}}[V(x_0, x)|x_t = x], \text{grad} \log \frac{\rho_t}{\nu} \right\rangle dx \tag{55}$$

$$= \int_M \rho_t \int_M \left\langle \text{grad} f(x) - V(x_0, x), \text{grad} \log \frac{\rho_t}{\nu} \right\rangle \rho_{0|t}(x_0|x) dx_0 dx \tag{56}$$

$$= \int_{M \times M} \rho_{0t}(x_0, x) \left\langle \text{grad} f(x) - V(x_0, x), \text{grad} \log \frac{\rho_t}{\nu} \right\rangle dx_0 dx \tag{57}$$

$$= \mathbb{E}_{\rho_{0t}} \left[\left\langle \text{grad} f(x) - V(x_0, x), \text{grad} \log \frac{\rho_t}{\nu} \right\rangle \right] \tag{58}$$

and then

$$\frac{d}{dt} H(\rho_t|\nu) = -\int_M \rho_t \left\| \text{grad} \log \frac{\rho_t}{\nu} \right\|^2 dx + \mathbb{E}_{\rho_{0t}} \left[\left\langle \text{grad} f(x) - V(x_0, x), \text{grad} \log \frac{\rho_t}{\nu} \right\rangle \right].$$

18

Moreover,

$$\mathbb{E}_{\rho_{0t}} \left[\left\langle \mathrm{grad} f(x) - V(x_0, x), \mathrm{grad} \log \frac{\rho_t}{\nu} \right\rangle \right] \tag{59}$$

$$= \mathbb{E}_{\rho_{0t}} \left[\left\langle \mathrm{grad} f(x) - \Gamma_{x_0}^x \mathrm{grad} f(x_0) + \Gamma_{x_0}^x \mathrm{grad} f(x_0) - V(x_0, x), \mathrm{grad} \log \frac{\rho_t}{\nu} \right\rangle \right] \tag{60}$$

$$= \mathbb{E}_{\rho_{0t}} \left[\left\langle \mathrm{grad} f(x) - \Gamma_{x_0}^x \mathrm{grad} f(x_0), \mathrm{grad} \log \frac{\rho_t}{\nu} \right\rangle \right] \tag{61}$$

$$+ \mathbb{E}_{\rho_{0t}} \left[\left\langle \Gamma_{x_0}^x \mathrm{grad} f(x_0) - V(x_0, x), \mathrm{grad} \log \frac{\rho_t}{\nu} \right\rangle \right] \tag{62}$$

$$\leq \mathbb{E}_{\rho_{0t}} \left[\| \mathrm{grad} f(x) - \Gamma_{x_0}^x \mathrm{grad} f(x_0) \|^2 \right] + \frac{1}{4} \mathbb{E}_{\rho_{0t}} \left[\left\| \mathrm{grad} \log \frac{\rho_t(x_t)}{\nu(x_t)} \right\|^2 \right] \tag{63}$$

$$+ \mathbb{E}_{\rho_{0t}} \left[\| \Gamma_{x_0}^x \mathrm{grad} f(x_0) - V(x_0, x) \|^2 \right] + \frac{1}{4} \mathbb{E}_{\rho_{0t}} \left[\left\| \mathrm{grad} \log \frac{\rho_t(x_t)}{\nu(x_t)} \right\|^2 \right] \tag{64}$$

$$= \mathbb{E}_{\rho_{0t}} \left[\| \mathrm{grad} f(x) - \Gamma_{x_0}^x \mathrm{grad} f(x_0) \|^2 \right] + \mathbb{E}_{\rho_{0t}} \left[\| \Gamma_{x_0}^x \mathrm{grad} f(x_0) - V(x_0, x) \|^2 \right] \tag{65}$$

$$+ \frac{1}{2} \mathbb{E}_{\rho_{0t}} \left[\left\| \mathrm{grad} \log \frac{\rho_t(x_t)}{\nu(x_t)} \right\|^2 \right] \tag{66}$$

By the Lipschitzness of gradient, i.e. $\| \mathrm{grad} f(x) - \Gamma_{x_0}^x \mathrm{grad} f(x_0) \| \leq L d(x_0, x)$, we have

$$\mathbb{E}_{\rho_{0t}} \left[\| \mathrm{grad} f(x) - \Gamma_{x_0}^x \mathrm{grad} f(x_0) \|^2 \right] \leq L^2 \mathbb{E}_{\rho_{0t}} [d(x, x_0)^2].$$

On the other hand, in the second term

$$\| V(x_0, x) - \Gamma_{x_0}^x \mathrm{grad} f(x_0) \| = \| D\mathrm{Exp}_{x_0}(\mathrm{Exp}_{x_0}^{-1}(x))[\mathrm{grad} f(x_0)] - \Gamma_{x_0}^x \mathrm{grad} f(x_0) \| \tag{67}$$

$$\leq \| D\mathrm{Exp}_{x_0}(\mathrm{Exp}_{x_0}^{-1}(x)) - \Gamma_{x_0}^x \|_{op} \| \mathrm{grad} f(x_0) \|. \tag{68}$$

Note that when $x = x_0$, $D\mathrm{Exp}_{x_0}(\mathrm{Exp}_{x_0}^{-1}(x)) = \Gamma_{x_0}^x = Id$ and the operator norm for $x \neq x_0$ depends on sectional curvature c of M. Let $v \in T_{x_0} M$, $\|v\| = 1$ and $\mathrm{Exp}_{x_0}(tv) = x$. Then

$$\left\| D\mathrm{Exp}_{x_0}(\mathrm{Exp}_{x_0}^{-1}(x)) - \Gamma_{x_0}^x \right\|_{op} = \left\| D\mathrm{Exp}_{x_0}(tv) - \Gamma_{x_0}^x \right\|_{op}$$

and $\left\| D\mathrm{Exp}_{x_0}(tv) - \Gamma_{x_0}^x \right\|_{op} \to 0$ as $t \to 0$, more specific, $\left\| D\mathrm{Exp}_{x_0}(tv) - \Gamma_{x_0}^x \right\|_{op} = O(t)$. Since M is compact, there exist a global constant $K(R) > 0$ determined by the Riemann curvature tensor, such that the square of the operator norm satisfies

$$\left\| D\mathrm{Exp}_{x_0}(tv) - \Gamma_{x_0}^x \right\|_{op}^2 \leq K(R)t^2.$$

And then

$$\mathbb{E}_{\rho_{0t}} \left[\| V(x_0, x) - \Gamma_{x_0}^x \mathrm{grad} f(x_0) \|^2 \right] \leq \left[\mathbb{E}_{\rho_{0t}} \left\| D\mathrm{Exp}_{x_0}(tv) - \Gamma_{x_0}^x \right\|_{op}^2 \| \mathrm{grad} f(x_0) \|^2 \right] \tag{69}$$

$$\leq K(R)t^2 \mathbb{E}_{\rho_0} \left[\| \mathrm{grad} f(x_0) \|^2 \right] \tag{70}$$

$$\leq K(R)t^2 \left(\frac{4L^2}{\alpha} H(\rho_0 | \nu) + 2nL \right). \tag{71}$$

Under normal coordinate systems, the Christoffel symbols Γ_{ij}^k can be chosen to be zero, the local Langevin algorithm can be written as

$$x_t = \exp_{x_0}(-t\nabla f(x_0) + \sqrt{2t} z), \; z \sim \mathcal{N}(0, I).$$

By exponential map, we have the distance $d(x_t, x_0)$ is the integral $\int_0^1 \left\| -t\nabla f(x_0) + \sqrt{2t} z \right\| ds$, so

$$d(x_t, x_0)^2 = \left\| -t\nabla f(x_0) + \sqrt{2t} z \right\|^2$$

and then

$$\mathbb{E}_{\rho_{0t}} \left[d(x_t, x_0)^2 \right] = \mathbb{E}_{\rho_{0t}} \left[\left\| -t\nabla f(x_0) + \sqrt{2t} z \right\|^2 \right] \tag{72}$$

$$= t^2 \mathbb{E}_{\rho_0} \left[\| \nabla f(x_0) \|^2 \right] + 2tn \tag{73}$$

$$\leq \frac{4t^2 L^2}{\alpha} H(\rho_0 | \nu) + 2t^2 nL + 2tn. \tag{74}$$

where the second equality is due to that z is independent of x_0 and the mean of $\langle z, z \rangle$ is n. So we have the discretization error bound

$$\mathbb{E}_{\rho_{0t}} \left[\left\langle \operatorname{grad} f(x) - V(x_0, x), \operatorname{grad} \log \frac{\rho_t}{\nu} \right\rangle \right] \leq L^2 \left(\frac{4t^2 L^2}{\alpha} H(\rho_0|\nu) + 2t^2 nL + 2tn \right) \tag{75}$$

$$+ K(R) t^2 \left(\frac{4L^2}{\alpha} H(\rho_0|\nu) + 2nL \right) \tag{76}$$

$$+ \frac{1}{2} \mathbb{E}_{\rho_{0t}} \left[\left\| \operatorname{grad} \log \frac{\rho_t(x_t)}{\nu(x_t)} \right\|^2 \right] \tag{77}$$

and the time derivative of KL divergence is bounded

$$\frac{d}{dt} H(\rho_t|\nu) \leq -\frac{1}{2} \mathbb{E}_{\rho_{0t}} \left[\left\| \operatorname{grad} \log \frac{\rho_t(x_t)}{\nu(x_t)} \right\|^2 \right] + L^2 \left(\frac{4t^2 L^2}{\alpha} H(\rho_0|\nu) + 2t^2 nL + 2tn \right) \tag{78}$$

$$+ K(R) t^2 \left(\frac{4L^2}{\alpha} H(\rho_0|\nu) + 2nL \right). \tag{79}$$

Denoting $K' = \frac{4L^2}{\alpha} H(\rho_0|\nu) + 2nL$ for short, by log-Sobolev inequality:

$$H(\rho_t|\nu) \leq \frac{1}{2\alpha} \mathbb{E}_{\rho_{0t}} \left[\left\| \operatorname{grad} \log \frac{\rho_t(x_t)}{\nu(x_t)} \right\|^2 \right]$$

we have

$$-\frac{1}{2} \mathbb{E}_{\rho_{0t}} \left[\left\| \operatorname{grad} \log \frac{\rho_t(x_t)}{\nu(x_t)} \right\|^2 \right] \leq -\alpha H(\rho_t|\nu)$$

and then

$$\frac{d}{dt} H(\rho_t|\nu) \leq -\alpha H(\rho_t|\nu) + L^2 \left(\frac{4t^2 L^2}{\alpha} H(\rho_0|\nu) + 2t^2 nL + 2tn \right) + K(R) K' t^2 \tag{80}$$

$$= -\alpha H(\rho_t|\nu) + \frac{4t^2 L^4}{\alpha} H(\rho_0|\nu) + 2t^2 nL^3 + 2tnL^2 + K(R) K' t^2 \tag{81}$$

$$= -\alpha H(\rho_t|\nu) + \frac{4t^2 L^2}{\alpha} (L^2 + K(R)) H(\rho_0|\nu) + 2t^2 nL^3 + 2tnL^2 + 2K(R) t^2 nL. \tag{82}$$

For $0 \leq t \leq \epsilon$, and taking $\epsilon \leq \frac{1}{2L}$, we have

$$\frac{d}{dt} H(\rho_t|\nu) \leq -\alpha H(\rho_t|\nu) + \frac{4\epsilon^2 L^2}{\alpha} (L^2 + K(R)) H(\rho_0|\nu) + 2\epsilon^2 nL^3 + 2\epsilon nL^2 + 2K(R) nL\epsilon^2 \tag{83}$$

$$\leq -\alpha H(\rho_t|\nu) + \frac{4\epsilon^2 L^2}{\alpha} (L^2 + K(R)) H(\rho_0|\nu) + 3\epsilon nL^2 + 2K(R) nL\epsilon^2, \tag{84}$$

Multiplying both sides by $e^{\alpha t}$, the following holds

$$e^{\alpha t} \frac{d}{dt} H(\rho_t|\nu) \leq e^{\alpha t} \left(-\alpha H(\rho_t|\nu) + \frac{4\epsilon^2 L^2}{\alpha} (L^2 + K(R)) H(\rho_0|\nu) + 3\epsilon nL^2 + 2K(R) nL\epsilon^2 \right)$$

and then

$$\frac{d}{dt} \left(e^{\alpha t} H(\rho_t|\nu) \right) = \alpha e^{\alpha t} H(\rho_t|\nu) + e^{\alpha t} \frac{d}{dt} H(\rho_t|\nu) \tag{85}$$

$$\leq \alpha e^{\alpha t} H(\rho_t|\nu) + e^{\alpha t} \left(-\alpha H(\rho_t|\nu) + \frac{4\epsilon^2 L^2}{\alpha} (L^2 + K(R)) H(\rho_0|\nu) + 3\epsilon nL^2 + 2K(R) nL\epsilon^2 \right) \tag{86}$$

$$= e^{\alpha t} \left(\frac{4\epsilon^2 L^2}{\alpha} (L^2 + K(R)) H(\rho_0|\nu) + 3\epsilon nL^2 + 2K(R) nL\epsilon^2 \right). \tag{87}$$

Integrating both sides for $t \in [0, \epsilon]$,

$$\int_0^\epsilon \frac{d}{dt} (e^{\alpha t} H(\rho_t|\nu)) dt \leq \int_0^\epsilon e^{\alpha t} \left(\frac{4\epsilon^2 L^2}{\alpha} (L^2 + K(R)) H(\rho_0|\nu) + 3\epsilon nL^2 + 2K(R) nL\epsilon^2 \right) dt \tag{88}$$

that implies

$$e^{\alpha \epsilon} H(\rho_\epsilon|\nu) - H(\rho_0|\nu) \leq \frac{e^{\alpha \epsilon} - 1}{\alpha} \left(\frac{4\epsilon^2 L^2}{\alpha} (L^2 + K(R)) H(\rho_0|\nu) + 3\epsilon nL^2 + 2K(R) nL\epsilon^2 \right).$$

Using inequality $e^{\alpha\epsilon} \leq 1 + 2\alpha\epsilon$ for $0 < \alpha\epsilon \leq 1$, i.e. $\epsilon \leq \frac{1}{\alpha}$, we have

$$e^{\alpha\epsilon}H(\rho_\epsilon|\nu) - H(\rho_0|\nu) \leq 2\epsilon\left(\frac{4\epsilon^2 L^2}{\alpha}(L^2 + K(R))H(\rho_0|\nu) + 3\epsilon nL^2 + 2K(R)nL\epsilon^2\right)$$

or equivalently

$$e^{\alpha\epsilon}H(\rho_\epsilon|\nu) \leq \left(1 + \frac{8\epsilon^3 L^2}{\alpha}(L^2 + K(R))\right)H(\rho_0|\nu) + 6\epsilon^2 nL^2 + 4K(R)nL\epsilon^3$$

and thus

$$H(\rho_\epsilon|\nu) \leq e^{-\alpha\epsilon}\left(1 + \frac{8\epsilon^3 L^2}{\alpha}(L^2 + K(R))\right)H(\rho_0|\nu) + e^{-\alpha\epsilon}(6\epsilon^2 nL^2 + 4K(R)nL\epsilon^3).$$

Choose ϵ small enough, i.e. $\epsilon \leq \frac{\alpha}{4L\sqrt{L^2+K(R)}}$ so that

$$1 + \frac{8\epsilon^3 L^2}{\alpha}(L^2 + K(R)) \leq 1 + \frac{\alpha\epsilon}{2} \leq e^{\frac{1}{2}\alpha\epsilon}$$

Since $e^{-\alpha\epsilon} \leq 1$, we have

$$H(\rho_\epsilon|\nu) \leq e^{-\frac{1}{2}\alpha\epsilon}H(\rho_0|\nu) + 6\epsilon^2 nL^2 + 4K(R)nL\epsilon^3$$

Renaming $\rho_{k+1} = \rho_\epsilon$ and $\rho_k = \rho_0$, the inequality becomes

$$H(\rho_{k+1}|\nu) \leq e^{-\frac{1}{2}\alpha\epsilon}H(\rho_k|\nu) + 6\epsilon^2 nL^2 + 4K(R)nL\epsilon^3$$

and the proof completes. □

Theorem 4.5. Suppose M is a compact manifold without boundary and R is the Riemann curvature, $\nu = e^{-f}$ a density on M with $\alpha > 0$ the log-Sobolev constant. Then there exists a global constant $K(R)$ determined by c, such that for any $x_0 \sim \rho_0$ with $H(\rho_0|\nu) \leq \infty$, the iterates $x_k \sim \rho_k$ of GLA with stepsize $\epsilon \leq \min\{\frac{\alpha}{4L\sqrt{L^2+K(R)}}, \frac{3L}{2K(R)}, \frac{1}{2L}, \frac{1}{2\alpha}\}$ satisfty

$$H(\rho_k|\nu) \leq e^{-\frac{k}{2}\alpha\epsilon}H(\rho_0|\nu) + \frac{32\epsilon nL^2}{\alpha}$$

Proof. By Lemma C.3, we have the following relation of $H(\rho_k|\nu)$ and $H(\rho_0|\nu)$,

$$H(\rho_k|\nu) \leq e^{-\frac{k}{2}\alpha\epsilon}H(\rho_0|\nu) + \frac{1}{1 - e^{-\frac{1}{2}\alpha\epsilon}}(6\epsilon^2 nL^2 + 4K(R)nL\epsilon^3). \tag{89}$$

Since for $0 < \alpha\epsilon \leq \frac{1}{2}$, (this also implies that $\epsilon \leq \frac{1}{\alpha}$), we have $1 - e^{-\frac{1}{2}\alpha\epsilon} \geq \frac{3}{8}\alpha\epsilon$, and then

$$H(\rho_k|\nu) \leq e^{-\frac{k}{2}\alpha\epsilon}H(\rho_0|\nu) + \frac{16\epsilon nL^2}{\alpha} + \frac{32K(R)nL\epsilon^2}{3\alpha}.$$

Choosing $\epsilon \leq \frac{3L}{2K(R)}$ so that $\frac{32K(R)nL\epsilon^2}{3\alpha} \leq \frac{16\epsilon nL^2}{\alpha}$, the above inequality can be simplified to

$$H(\rho_k|\nu) \leq e^{-\frac{k}{2}\alpha\epsilon}H(\rho_0|\nu) + \frac{32\epsilon nL^2}{\alpha}$$

□

C.3   Proof of Theorem 4.6

Next we prove the rate of convergence for constant curvature, where the key technique is using Jacobi field that decompose any vector field into the tangential vector field and the normal vector field. Given a regular curve $\gamma : I \to M$, for each $t \in I$ we let $T_{\gamma(t)}^\top M$ denote the one-dimensional subspace spanned by $\gamma'(t)$, and $T_{\gamma t}^\perp M$ its orthogonal complement. A tangential vector field along γ is a vector field $V(t)$ such that $V(t) \in T_{\gamma(t)}^\top M$ for all t, and a normal vector field along γ is one such that $V(t) \in T_{\gamma t\perp} M$ for all t. A normal Jacobi field along γ is a Jacobi field $J(t)$ satisfying $J(t) \perp \gamma'(t)$ for all t.

Theorem 4.6. Suppose M is a compact manifold without boundary and c is the constant sectional curvature, $\nu = e^{-f}$ a density on M with $\alpha > 0$ the log-Sobolev constant. Then for stepsize $\epsilon \leq \min\{\frac{\sqrt{36}L^2}{|c|}, \frac{1}{2L}, \frac{1}{2\alpha}, \frac{\alpha}{\sqrt{32L^2}}\}$, for any $x_0 \sim \rho_0$ with $H(\rho_0|\nu) \leq \infty$, the iterates $x_k \sim \rho_k$ of GLA satisfy

$$H(\rho_k|\nu) \leq e^{-\frac{k}{2}\alpha\epsilon}H(\rho_0|\nu) + \frac{64\epsilon nL^2}{3\alpha}.$$

Proof. The main strategy of the proof agrees with proof of Lemma C.3. We recall some setting up of notations and results presented in previous proofs. Given a point x_0 on M, we denote $V(x_0, x)$ the vector defined by

$$V(x_0, x) = D\mathrm{Exp}_{x_0}\left(\mathrm{Exp}_{x_0}^{-1}(x)\right)[\mathrm{grad} f(x_0)],$$

and ρ_{0t} the joint distribution satisfying

$$\rho_{0t}(x_0, x_t) = \rho_0(x_0)\rho_{t|0}(x_t, x_0) = \rho_t(x_t)\rho_{0|t}(x_0, x_t).$$

From the proof of Lemma C.3, we have

$$\frac{d}{dt}H(\rho_t|\nu) = -\int_M \rho_t \left\|\mathrm{grad}\log\frac{\rho_t}{\nu}\right\|^2 dx + \mathbb{E}_{\rho_{0t}}\left[\left\langle\mathrm{grad} f(x) - V(x_0, x), \mathrm{grad}\log\frac{\rho_t}{\nu}\right\rangle\right].$$

where

$$\mathbb{E}_{\rho_{0t}}\left[\left\langle\mathrm{grad} f(x) - V(x_0, x), \mathrm{grad}\log\frac{\rho_t}{\nu}\right\rangle\right] \leq \mathbb{E}_{\rho_{0t}}\left[\|\mathrm{grad} f(x) - \Gamma_{x_0}^x \mathrm{grad} f(x_0)\|^2\right] \tag{90}$$

$$+ \mathbb{E}_{\rho_{0t}}\left[\|\Gamma_{x_0}^x \mathrm{grad} f(x_0) - V(x_0, x)\|^2\right] \tag{91}$$

$$+ \frac{1}{2}\mathbb{E}_{\rho_{0t}}\left[\left\|\mathrm{grad}\log\frac{\rho_t(x_t)}{\nu(x_t)}\right\|^2\right]. \tag{92}$$

By the Lipschitzness of gradient, $\|\mathrm{grad} f(x) - \Gamma_{x_0}^x \mathrm{grad} f(x_0)\| \leq Ld(x_0, x)$, it holds that

$$\mathbb{E}_{\rho_{0t}}\left[\|\mathrm{grad} f(x) - \Gamma_{x_0}^x \mathrm{grad} f(x_0)\|^2\right] \leq L^2 \mathbb{E}_{\rho_{0t}}[d(x_0, x)^2].$$

So the difference from Lemma C.3 lies in the bound of $\mathbb{E}_{\rho_{0t}}\left[\|\Gamma_{x_0}^x \mathrm{grad} f(x_0) - V(x_0, x)\|^2\right]$ in the setting of constant sectional curvature. We complete this by following standard arguments using Jacobi field, e.g. Lemma E.2 in Agarwal et al. [2020]. Given any two tangent vectors $v, s \in T_{x_0}$, $J(t) = D\mathrm{Exp}_{x_0}(tv)[ts]$ is the unique Jacobi field along the geodesic $\gamma(t) = \mathrm{Exp}_{x_0}(tv)$ satisfying $J(0) = 0$ and $\frac{D}{dt}J(0) = s$. If $s = rv$ for some $r \in \mathbb{R}$ so that s and v are parallel, then

$$J(t) = D\mathrm{Exp}_{x_0}(tv)[ts] = \frac{d}{dq}\mathrm{Exp}_{x_0}(tv + qts)\Big|_{q=0} = \frac{d}{dq}\gamma(t + qrt)\Big|_{q=0} = rt\gamma'(t) = tP_{tv}(s),$$

where we simplify the notation in that $P_{tv} = \Gamma_{x_0}^{\mathrm{Exp}_{x_0}(tv)}$ is the parallel transport along the geodesic $\gamma(t)$ connecting x_0 and x with initial velocity v.

If M has constant sectional curvature c, $\|v\| = 1$ and $\langle v, s\rangle = 0$, the Jacobi field can be written as

$$J(t) = s_c(t)P_{tv}(s),$$

where

$$s_c(t) = \begin{cases} t & \text{if } c = 0, \\ \frac{\sin(\sqrt{c}t)}{\sqrt{c}} & \text{if } c > 0, \\ \frac{\sinh(\sqrt{-c}t)}{\sqrt{-c}} & \text{if } c < 0. \end{cases}$$

Evaluating at $t = 1$ and by linearity in s, we have that

$$D\mathrm{Exp}_{x_0}(v)[s] = P_v\left(s_\| + \frac{s_c(\|v\|)}{\|v\|}s_\perp\right)$$

and then, by definition of P_v, the following holds,

$$D\mathrm{Exp}_{x_0}(v)[s] - \Gamma_{x_0}^x s = D\mathrm{Exp}_{x_0}(v)[s] - P_v(s) \tag{93}$$

$$= P_v\left(s_\| + \frac{s_c(\|v\|)}{\|v\|}s_\perp\right) - P_v\left(s_\| + s_\perp\right) \tag{94}$$

$$= P_v\left(\frac{s_c(\|v\|)}{\|v\|}s_\perp - s_\perp\right) \tag{95}$$

$$= \left(\frac{s_c(\|v\|)}{\|v\|} - 1\right)P_v(s_\perp). \tag{96}$$

Since parallel transport is isometry and $\|s_\perp\| \leq \|s\|$, the norm of above vector satisfies

$$\left\|D\mathrm{Exp}_{x_0}(v)[s] - \Gamma_{x_0}^x s\right\| = \left|\frac{s_c(\|v\|)}{\|v\|} - 1\right|\|P_v(s_\perp)\| \tag{97}$$

$$= \left|\frac{s_c(\|v\|)}{\|v\|} - 1\right|\|s_\perp\| \tag{98}$$

$$\leq \left|\frac{s_c(\|v\|)}{\|v\|} - 1\right|\|s\|. \tag{99}$$

22

Letting $s = \mathrm{grad}f(x_0)$ and $\|tv\| = t$ for unit vector v, we have

$$\left\|D\mathrm{Exp}_{x_0}(\mathrm{Exp}_{x_0}^{-1}(x)) - \Gamma_{x_0}^{x}\mathrm{grad}f(x_0)\right\|^2 \leq h^2(t)\left\|\mathrm{grad}f(x_0)\right\|^2$$

where

$$h(t) = \begin{cases} 0 & \text{if } c = 0, \\ \frac{\sin(\sqrt{c}t)}{\sqrt{c}t} - 1 & \text{if } c > 0, \\ \frac{\sinh(\sqrt{-c}t)}{\sqrt{-c}t} - 1 & \text{if } c < 0. \end{cases}$$

Furthermore, the function $|h(t)|$ behaves like $\frac{c}{6}t^2$ for small t, independent of the position of x_0.

Repeating the arguments as in the proof of Lemma C.3, the time derivative of KL divergence is bounded by

$$\frac{d}{dt}H(\rho_t|\nu) \leq -\frac{1}{2}\mathbb{E}_{\rho_{0t}}\left[\left\|\mathrm{grad}\log\frac{\rho_t(x_t)}{\nu(x_t)}\right\|^2\right] + L^2\left(\frac{4t^2L^2}{\alpha}H(\rho_0|\nu) + 2t^2nL + 2tn\right) \tag{100}$$

$$+ h^2(t)\left(\frac{4L^2}{\alpha}H(\rho_0|\nu) + 2nL\right) \tag{101}$$

$$= -\frac{1}{2}\mathbb{E}_{\rho_{0t}}\left[\left\|\mathrm{grad}\log\frac{\rho_t(x_t)}{\nu(x_t)}\right\|^2\right] + \frac{4L^2}{\alpha}\left(L^2t^2 + h^2(t)\right)H(\rho_0|\nu) \tag{102}$$

$$+ 2t^2nL^3 + 2tnL^2 + 2h^2(t)nL. \tag{103}$$

When t is small and by choosing t such that $t^2 \leq \frac{36L^2}{c^2}$, the inequality can be simplified as

$$\frac{d}{dt}H(\rho_t|\nu) \leq -\frac{1}{2}\mathbb{E}_{\rho_{0t}}\left[\left\|\mathrm{grad}\log\frac{\rho_t(x_t)}{\nu(x_t)}\right\|^2\right] + \frac{8L^4t^2}{\alpha}H(\rho_0|\nu) \tag{104}$$

$$+ 4t^2nL^3 + 2tnL^2 \tag{105}$$

and by choosing $\epsilon \leq \frac{1}{2L}$ and LSI, we conclude that

$$\frac{d}{dt}H(\rho_t|\nu) \leq -\alpha H(\rho_t|\nu) + \frac{8L^4\epsilon^2}{\alpha}H(\rho_0|\nu) + 4\epsilon nL^2.$$

Multiplying both sides by $e^{\alpha t}$ and rearranging, we have

$$\frac{d}{dt}(e^{\alpha t}H(\rho_t|\nu)) \leq e^{\alpha t}\left(\frac{8L^4\epsilon^2}{\alpha}H(\rho_0|\nu) + 4\epsilon nL^2\right).$$

Integrating t from 0 to ϵ,

$$\int_0^\epsilon \frac{d}{dt}(e^{\alpha t}H(\rho_t|\nu))dt = e^{\alpha\epsilon}H(\rho_\epsilon|\nu) - H(\rho_0|\nu) \leq \frac{e^{\alpha\epsilon}-1}{\alpha}\left(\frac{8L^4\epsilon^2}{\alpha}H(\rho_0|\nu) + 4\epsilon nL^2\right)$$

By choosing ϵ small such that $e^{\alpha\epsilon} - 1 \leq 2\alpha\epsilon$ and then

$$e^{\alpha\epsilon}H(\rho_\epsilon|\nu) \leq H(\rho_0|\nu) + 2\epsilon\left(\frac{8L^4\epsilon^2}{\alpha}H(\rho_0|\nu) + 4\epsilon nL^2\right) \tag{106}$$

$$= \left(1 + \frac{16L^4\epsilon^3}{\alpha}\right)H(\rho_0|\nu) + 8\epsilon^2nL^2. \tag{107}$$

Let $\epsilon < \frac{\alpha}{\sqrt{32L^2}}$ such that $1 + \frac{16L^4\epsilon^3}{\alpha} \leq 1 + \frac{\alpha\epsilon}{2} \leq e^{\frac{1}{2}\alpha\epsilon}$, and then the following inequality holds

$$H(\rho_\epsilon|\nu) \leq e^{-\frac{1}{2}\alpha\epsilon}H(\rho_0|\nu) + 8\epsilon^2nL^2.$$

Renaming $\rho_0 = \rho_k$ and $\rho_\epsilon = \rho_{k+1}$, we have

$$H(\rho_{k+1}|\nu) \leq e^{-\frac{1}{2}\alpha\epsilon}H(\rho_k|\nu) + 8\epsilon^2nL^2$$

and apply this relation recurrently, we can prove

$$H(\rho_k|\nu) \leq e^{-\frac{k}{2}\alpha\epsilon}H(\rho_0|\nu) + \frac{1}{1-e^{-\frac{1}{2}\alpha\epsilon}}(8\epsilon^2nL^2)$$

Since $1 - e^{-\frac{1}{2}\alpha\epsilon} \geq \frac{3}{8}\alpha\epsilon$ for $\alpha\epsilon < \frac{1}{2}$, the above inequality can be written as

$$H(\rho_k|\nu) \leq e^{-\frac{k}{2}\alpha\epsilon}H(\rho_0|\nu) + \frac{64\epsilon nL^2}{3\alpha}$$

$\square$

D Experiments

As mentioned before, for simplicity, we can implement GLA without using the exponential map where a geodesic ODE solver is required, especially for the case when M is a submanifold of $\mathbb{R}^n$. In general, the retraction map from $T_x M$ to M is used in optimization on Riemannian manifold Agarwal et al. [2020], as a replacement of exponential map. In this section, we give experiments on sampling from distributions on the unit sphere in comparison of exponential map and orthogonal projection as a retraction in the geodesic step of GLA.

The experiments are designed to verify the following properties:

1. GLA captures the target distribution e^{-f} as expected;

2. The projection map behaves well in replacing the exponential map without solving geodesic equations.

In each set of figures, (a) is the landscape of the ideal distribution, (b) and (c) are the results with small number of iterations for exponential map and projection, (d) and (e) are enhanced with large number of iterations. We start with the definition of the general retraction in optimization on manifold.

Definition D.1 (Retraction). A retraction on a manifold M is a smooth mapping Retr from the tangent bundle TM to M satisfying properties 1 and 2 below: Let $\mathrm{Retr}_x : T_x M \to M$ denote the restriction of Retr to $T_x M$.

1. $\mathrm{Retr}_x(0) = x$, where 0 is the zero vector in $T_x M$.

2. The differential of Retr_x at 0 is the identity map.

Suppose M is a submanifold of $\mathbb{R}^n$ with positive codimension. Denote $\mathrm{Proj}_{T_x M}$ the orthogonal projection to the tangent space at x, then the retraction can be defined as $\mathrm{Retr}_x(v) = \mathrm{Proj}_M(x+v)$. The GLA on a submanifold of $\mathbb{R}^n$ can be written as

$$x_{k+1} = \mathrm{Retr}_x\left(\mathrm{Proj}_{T_x M}(-\epsilon \nabla f(x_k) + \sqrt{2\epsilon}\xi_0)\right) \tag{108}$$
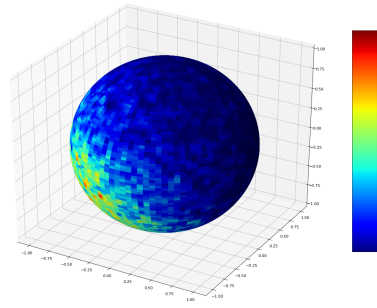
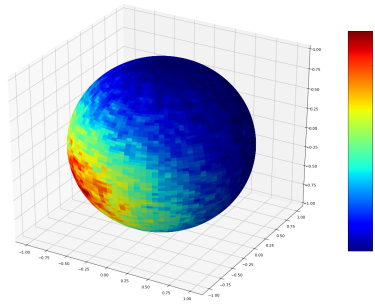If $M = S^{n-1}$ be the unit sphere in $\mathbb{R}^n$, then $\mathrm{Retr}_x(v) = \frac{x+v}{\|x+v\|}$.
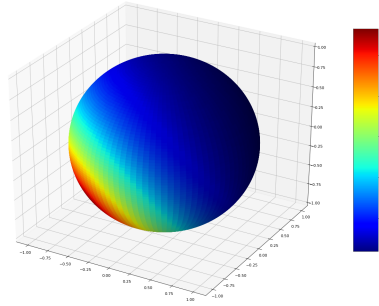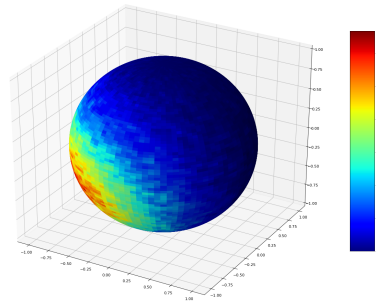
(a) Ideal distribution of e^{-f}
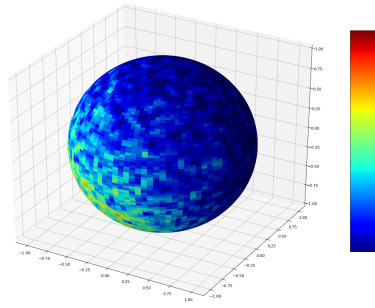
(b) $\mathrm{Exp}_x(v)$, iteration: 10k
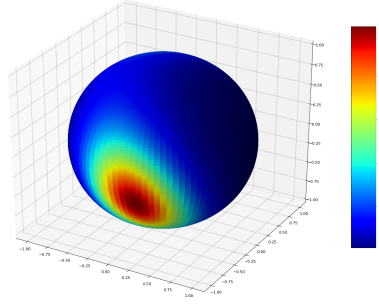

(c) $\mathrm{Retr}_x(v)$, iteration: 10k
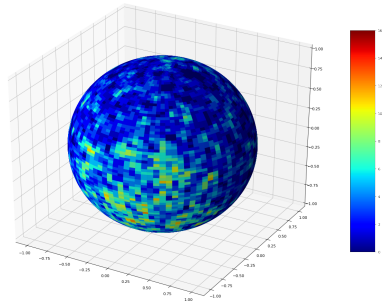

(d) $\mathrm{Exp}_x(v)$, iteration: 100k
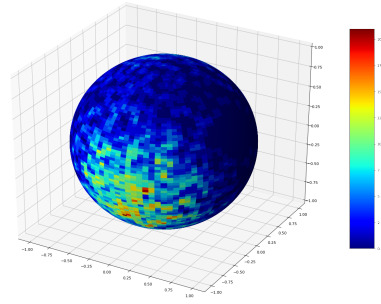

(e) $\mathrm{Retr}_x(v)$, iteration: 100k

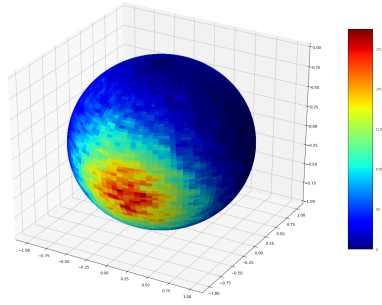Figure 2: $f(x) = x_1 + x_2 + x_3$, stepsize $\epsilon = 0.1$.

(a) Ideal distribution of e^{-f}

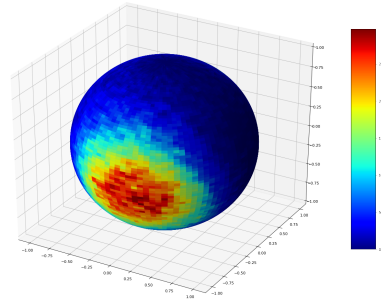(b) $\mathrm{Exp}_x(v)$, iteration: 10k

(c) $\mathrm{Retr}_x(v)$, iteration: 10k

(d) $\mathrm{Exp}_x(v)$, iteration: 100k

(e) $\mathrm{Retr}_x(v)$, iteration: 100k

Figure 3: $f(x) = x_1^2 + 3.05x_2^2 - 0.9x_3^2 + 1.1x_1x_2 + -1.02x_2x_3 + 2.1x_3x_1$, setpsize $\epsilon = 0.1$.