Riemannian Proximal Sampler for High-accuracy Sampling on Manifolds

Yunrui Guan
Department of Computational Applied Mathematics and Operations Research
Rice University
Houston, TX 77005
yg83@rice.edu

Krishnakumar Balasubramanian
Department of Statistics
University of California, Davis
Davis, CA 95616
kbala@ucdavis.edu

Shiqian Ma
Department of Computational Applied Mathematics and Operations Research
Rice University
Houston, TX 77005
sqma@rice.edu

Abstract

We introduce the *Riemannian Proximal Sampler*, a method for sampling from densities defined on Riemannian manifolds. The performance of this sampler critically depends on two key oracles: the *Manifold Brownian Increments (MBI)* oracle and the *Riemannian Heat-kernel (RHK)* oracle. We establish high-accuracy sampling guarantees for the Riemannian Proximal Sampler, showing that generating samples with ε-accuracy requires $\mathcal{O}(\log(1/\varepsilon))$ iterations in Kullback-Leibler divergence assuming access to exact oracles and $\mathcal{O}(\log^2(1/\varepsilon))$ iterations in the total variation metric assuming access to sufficiently accurate inexact oracles. Furthermore, we present practical implementations of these oracles by leveraging heat-kernel truncation and Varadhan's asymptotics. In the latter case, we interpret the Riemannian Proximal Sampler as a discretization of the entropy-regularized Riemannian Proximal Point Method on the associated Wasserstein space. We provide preliminary numerical results that illustrate the effectiveness of the proposed methodology.

1 Introduction

We study the problem of sampling from a density $\pi^X \propto e^{-f}$ on a Riemannian manifold (M, g), where g is the metric and π^X is defined with respect to the volume measure dV_g. The normalization constant $\int_M e^{-f}\, dV_g$ is unknown. Riemannian sampling arises in Bayesian inference (e.g., hierarchical models and Bayesian deep learning) [Girolami and Calderhead, 2011, Byrne and Girolami, 2013, Patterson and Teh, 2013, Liu and Zhu, 2018, Arnaudon et al., 2019, Liu et al., 2016, Piggott and Solo, 2016, Muniz et al., 2022, Lie et al., 2023], statistical physics (e.g., constrained molecular dynamics) [Leimkuhler and Matthews, 2016], and manifold optimization problems such as eigenvalue computation, low-rank approximation, and diffusion models [Goyal and Shetty, 2019, Li and Erdogdu, 2023, Yu et al., 2023, Bonet et al., 2023, De Bortoli et al., 2022, Huang et al., 2022].

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Table 1: Comparison of iteration complexity with Li and Erdogdu [2023] and Cheng et al. [2022]. Here, ε is the target accuracy, K_2 is the gradient Lipschitz constant, and d is the manifold dimension; for a product of n spheres of dimension m, $d = mn$. All works assume a uniform lower bound on Ricci curvature and include the hypersphere as a common example. Cheng et al. [2022] does not provide explicit dependence on d or other problem parameters.

Assumptions	Source	Complexity	Metric
LSI, $M = \mathcal{S}^m \times \mathcal{S}^m \times ... \times \mathcal{S}^m$ f is K_2-smooth	Li and Erdogdu [2023]	$\mathcal{O}\left(\frac{dK_2^2}{\alpha^2\varepsilon} \log \frac{H_{\pi^X}(\rho_0)}{\varepsilon}\right)$	KL
Distant-dissipativity, f isK_2-smooth, M has bounded sectional curvature	Cheng et al. [2022]	$\mathcal{O}\left(\frac{1}{\varepsilon^2} \operatorname{Poly}(d, K_2)\right)$	W_1
LSI, $M = \mathcal{S}^d$, f is L_1-Lipschitz	Corollary 10	$\tilde{\mathcal{O}}(\frac{L_1^2 d}{\alpha} \log^2 \frac{1}{\varepsilon})$	TV

On a Riemannian manifold, Langevin dynamics takes the form $dX_t = -\operatorname{grad} f(X_t)\, dt + \sqrt{2}\, dB_t$, where grad is the Riemannian gradient and B_t is Brownian motion on the manifold. This formulation extends Euclidean Langevin dynamics by incorporating geometric information through the Riemannian metric, but discretizing manifold Brownian motion is generally intractable. Li and Erdogdu [2023] considered product manifolds $\mathcal{S}^m \times \cdots \times \mathcal{S}^m$, showing convergence for a scheme that discretizes only the drift while assuming exact Brownian motion—feasible on spheres. Gatmiry and Vempala [2022] extended this to general Hessian manifolds, however, requiring exact Brownian motion. Both assume a log-Sobolev inequality and achieve poly$(1/\varepsilon)$ iteration complexity in KL divergence, but the need for exact Brownian motion in Gatmiry and Vempala [2022] limits practical applicability.

Cheng et al. [2022] analyzed a practical discretization of Riemannian Langevin dynamics, where both drift and noise are discretized. They established $\tilde{\mathcal{O}}(1/\varepsilon^2)$ iteration complexity in the 1-Wasserstein distance under general assumptions, and in the 2-Wasserstein distance under a stronger, log-concavity-like condition. A key challenge is proving Wasserstein contractivity without convexity (e.g., on compact manifolds), addressed via a second-order expansion of the Jacobi equation [Cheng et al., 2022, Lemma 29]. Kong and Tao [2024] proposed a Lie-group MCMC sampler for densities on Lie groups, achieving polynomial $\tilde{\mathcal{O}}(\text{poly}(1/\varepsilon))$ complexity in 2-Wasserstein distance.

In comparison to the above works for Riemannian sampling, for the Euclidean case, high-accuracy algorithms, i.e., algorithms with iteration complexity of $\tilde{\mathcal{O}}(\text{polylog}(1/\varepsilon))$ are available under various assumptions (that are essentially based on (strong) log-concavity or isoperimetry); see for example Lee et al. [2021], Chen et al. [2022], Fan et al. [2023], He et al. [2024] for such results for the Euclidean proximal sampler and Dwivedi et al. [2019], Chen et al. [2020], Chewi et al. [2021], Lee et al. [2020], Wu et al. [2022], Chen and Gatmiry [2023], Andrieu et al. [2024], Altschuler and Chewi [2024] for various Metropolized algorithms including Metropolis Random Walk (MRW), Metropolis Adjusted Langevin Algorithm (MALA) and Metropolis Hamiltonian Monte Carlo (MHMC).

High-accuracy samplers for constrained Euclidean sampling—i.e., from densities supported on convex sets $\mathcal{K} \subseteq \mathbb{R}^d$—have been developed using Hit-and-Run and Ball Walk under various conditions [Lovász, 1999, Kannan et al., 2006, 1997]; see Kook and Zhang [2025, Section 1.3] for a survey. Kook et al. [2022] introduced Constrained Riemannian HMC (CRHMC) with an Implicit Midpoint integrator and proved high-accuracy guarantees. Noble et al. [2023] proposed Barrier HMC (BHMC) and its discretizations with asymptotic guarantees. Kook et al. [2024] developed the "In-and-Out" algorithm for uniform sampling on convex bodies. Kook and Vempala [2024] achieved state-of-the-art accuracy for log-concave sampling via a proximal method. Srinivasan et al. [2024a,b] showed that Metropolized Mirror and preconditioned Langevin samplers also achieve high-accuracy under suitable assumptions.

Given the above, the following natural question arises:

Can one develop high-accuracy algorithms for sampling on Riemannian manifolds?

To the best of our knowledge, no prior work exists on providing an affirmative answer to this question. In this work, we develop the *Riemannian Proximal Sampler* which generalizes the Euclidean Proximal Sampler from Lee et al. [2021]. In contrast to the Euclidean case, the algorithm is based on the availability of two oracles: the Manifold Brownian Increment (MBI) oracle and the Riemannian Heat Kernel (RHK) oracle. We show in Theorem 6 and Theorem 8 that the algorithm achieves high-accuracy guarantees under functional inequality assumptions when exact oracles are available, and under Assumption 1 when inexact oracles are available, respectively. We further develop practical implementations of the aforementioned oracles that satisfy the conditions in Assumption 1 (Section 5), and that are connected to entropy-regularized proximal point method on Wasserstein spaces (Appendix A). A comparison with the existing results for the case of sampling on spheres is provided in Table 1. We also demonstrate the numerical performance of the algorithms via simulations in Appendix B.

2 Preliminaries

Throughout the paper, unless otherwise specified, we use $\tilde{\mathcal{O}}$, to suppress dependency on other parameters except for ε, and only keep leading factor. For example, $\frac{1}{\alpha}\log(\frac{1}{\varepsilon})(\log\log\frac{1}{\varepsilon}) = \tilde{\mathcal{O}}(\log\frac{1}{\varepsilon})$. We first recall certain preliminaries on Riemannian manifolds; additional preliminaries are provided in Appendix C. We refer the readers to Lee [2018] for more details.

Let M be a Riemannian manifold of dimension d equipped with metric g. The manifold M is assumed to be complete, connected Riemannian manifold without boundary. For a point $x \in M$, T_xM denotes the tangent space at x. For any $v, w \in T_xM$, we can write the metric as $g_x(v,w) = \langle v, w\rangle_g$. For $x \in M$ and $v \in T_xM$, $\exp_x(v)$ denotes the exponential map. We use grad and dV_g to represent the Riemannian gradient and the Riemannian volume form respectively.

For $x \in M$, $\mathrm{Cut}\,(x)$ denotes the cut locus of x. For $x, y \in M$, we use $d(x,y)$ to denote the geodesic distance between x and y. Let div denotes the Riemannian divergence, and Laplace-Beltrami operator $\Delta : C^\infty(M) \to C^\infty(M)$ is defined as the Riemannian divergence of Riemannian gradient: $\Delta u = \mathrm{div}\,(\mathrm{grad}\ u)$. We use $\nu(t,x,y)$ to denote the density of manifold Brownian motion with time t, starting at x, evaluated at y.

Let $(M, \mathcal{F})$ be a measurable space. Note that the Riemannian volume form dV_g is a measure. A probability measure ρ and its corresponding probability density function p are related through $d\rho = pdV_g$. Given a measurable set $A \in \mathcal{F}$, $P_\rho(A)$ denotes the probability assigned to the set A by ρ. We have $P_\rho(A) = \int_A p(x)dV_g(x) = \int_A d\rho(x)$.

Definition 1 (TV distance)**.** *Let ρ_1, ρ_2 be probability measures defined on the measurable space $(M, \mathcal{F})$. The total variation distance between ρ_1 and ρ_2 is defined as $\|\rho_1 - \rho_2\|_{TV} := \sup_{A\in\mathcal{F}} |\rho_1(A) - \rho_2(A)|$.*

Definition 2 (KL divergence and χ^2 divergence)**.** *Let ρ_1, ρ_2 be probability measures on the measurable space $(M, \mathcal{F})$, with full support. The Kullback-Leibler (KL) divergence and χ^2 divergence of ρ_1 with respect to ρ_2 are defined as (respectively)*

$$H_{\rho_2}(\rho_1) := \int_M \log\frac{d\rho_1}{d\rho_2}d\rho_1, \quad \chi^2_{\rho_2}(\rho_1) = \int_M (\frac{d\rho_1}{d\rho_2} - 1)^2 d\rho_2$$

where $\frac{d\rho_1}{d\rho_2}$ is the Radon-Nikodym derivative.

It is known that $H_{\rho_2}(\rho_1) \geq 0$ with equality if and only if $\rho_1 = \rho_2$. Although the KL divergence is not symmetric, it serve as a "distance" function between two probability measures. For instance, the well known Pinsker inequality states that $\|\rho_2 - \rho_1\|^2_{TV} \leq \frac{1}{2}H_{\rho_2}(\rho_1)$.

Definition 3 (Log-Sobolev Inequality (LSI))**.** *A probability measure ρ_2 satisfies Log-Sobolev Inequality with parameter $\alpha > 0$ (α-LSI) if $H_{\rho_2}(\rho_1) \leq \frac{1}{2\alpha}J_{\rho_2}(\rho_1), \forall\rho_1$, where $J_{\rho_2}(\rho_1) := \int_M \|\mathrm{grad}\log\frac{\rho_1}{\rho_2}\|^2 d\rho_1$ is the relative Fisher information.*

We also recall the definition of Poincaré inequality which is a generalization of LSI.

Definition 4 (Poincaré Inequality (PI))**.** *A probability measure ρ satisfies Poincaré Inequality with parameter $\alpha > 0$ (α-PI) if $\mathbb{E}_\rho(g^2) - \mathbb{E}_\rho[g]^2 \leq \frac{1}{\alpha}\mathbb{E}_\rho[\|\mathrm{grad}\ g\|^2], \forall g \in C^\infty(M)$*

For more technical details on LSI and PI, see Appendix H.2 and H.3. In Euclidean space, conditions like LSI and PI can be viewed as a relaxation of strong convexity assumption on f, and is used to establish convergence of sampling algorithms in KL divergence. See, for example, Vempala and Wibisono [2019] (for the Langevin Monte Carlo Algorithm) and Chen et al. [2022] (for the Euclidean proximal sampler). For a Riemannian manifold, the Bakry-Émery condition can be used to establish LSI. Informally speaking, when the potential f satisfies certain convexity, the corresponding probability measure satisfies LSI. For more details see for example Bakry et al. [2014] and [Li and Erdogdu, 2023, Appendix B]. When the manifold is compact, it is well known that the only convex function is the constant function, and therefore the Bakry-Émery condition does not yield useful information; but for non-compact manifolds, such a condition may serve as a useful tool to establish LSI. Moreover, recent works translated LSI/PI conditions on π to the Polyak-Lojasiewicz (PL) condition on f. For example, considering $e^{-f(x)/t}$ in the low-temperature regime, i.e., $t \to 0$ limit, Chewi and Stromme [2024] related LSI constant and the PL constant. Similarly Gong et al. [2024] related the PI constant and a local PL constant. Chen and Sridharan [2024] considered an "optimizability" condition and analyzed LSI/PI constant for (informally) $t \leq O(1/d)$. It is interesting future work to establish similar relationships in the manifold setting.

2.1 Curvature

We also need notions of curvature on manifolds to present our main results. Let $\mathfrak{X}(M)$ denote the set of all smooth vector fields on M. Define a map called Riemann curvature endomorphism by $R : \mathfrak{X}(M) \times \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M)$ by $R(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z$. While such definition is very abstract, we provide an intuitive explanation of what curvature is. Intuitively, on a manifold of positive curvature (say, a 2-dimensional sphere), geodesics tend to "contract". More precisely, given $x, y \in M$ and $v \in T_x M$, we can parallel transport v to $u = P_x^y v \in T_y M$. It is a well-known result that (ignore higher order terms) $d(\exp_x tv, \exp_y tu) \leq (1 - \frac{t^2}{2}K) d(x,y)$ for some K (which is actually the sectional curvature). From this, we see that for positive curvature, which means $K > 0$, the distance between geodesics would decrease.

Formally, given $v, w \in T_p M$ being linearly independent, the sectional curvature of the plane spanned by v and w can be computed through $K(v,w) = \frac{\langle R(v,w)w, v\rangle}{|v|^2|w|^2 - \langle v,w\rangle^2}$; see Lee [2018, Proposition 8.29]. On the other hand, Ricci curvature can be viewed as the average of sectional curvatures. The Ricci curvature at $x \in M$ along direction v is denoted as $\mathrm{Ric}_x(v)$, which is equal to the sum of the sectional curvatures of the 2-planes spanned by $(v, b_i)_{i=2}^d$ where $v, b_2, ..., b_d$ is an orthonormal basis for $T_x M$; see Lee [2018, Proposition 8.32].

We remark that the Ricci curvature is actually a symmetric 2-tensor field defined as the trace of the curvature endomorphism on its first and last indices [Lee, 2018], which sometimes is written as $\mathrm{Ric}_x(u,v)$ for $u, v \in T_x M$. The previous notation is a shorthand of $\mathrm{Ric}_x(v) = \mathrm{Ric}_x(v,v)$. When we say Ricci curvature is lower bounded by κ, we mean $\mathrm{Ric}(v,v) \geq \kappa, \forall v \in T_x M, \|v\| = 1$. We end this subsection through some concrete examples.

1. The hypersphere $\mathcal{S}^d$ has constant sectional curvature equal to 1, and constant Ricci curvature $\mathrm{Ric} = (d-1)g, \forall x \in M$ (so that $\mathrm{Ric}_x(v) = d-1$ for all unit tangent vector $v \in T_x M$).
2. For $P_m \subseteq \mathbb{R}^{m \times m}$, the manifold of positive definite matrices (with affine-invariant metric), its sectional curvatures are in the interval $[-\frac{1}{2}, 0]$; see, for example, Criscitiello and Boumal [2023]. Hence its Ricci curvature is lower bounded by $-\frac{m(m+1)-1}{4}$.

2.2 Brownian motion on manifolds

Now we briefly discuss Brownian motion on a Riemannian manifold. Recall that in Euclidean space, Brownian motion is described by the Wiener process. Given $x \in \mathbb{R}^d$ and $t > 0$, the Brownian motion starting at x with time t has (a Gaussian) density function $\nu(t,x,y) = \frac{1}{(2\pi t)^{d/2}} e^{-\frac{\|x-y\|^2}{2t}}$. It solves the heat equation $\frac{\partial}{\partial t}\nu(t,x,y) = \frac{1}{2}\Delta_y \nu(t,x,y)$ with initial condition $\nu(0,x,y) = \delta_x(y)$.

On a Riemannian manifold, we can describe the density of Brownian motion (heat kernel) through heat equation. Let $B_{x,t}$ be a random variable denoting manifold Brownian motion starting at x with time t and let $\nu(t,x,y)$ be the density of $B_{x,t}$. The Brownian motion density $\nu(t,x,y)$ is then defined

Algorithm 1 Riemannian proximal sampler

for $k = 0, 1, 2, ...$ **do**

Step 1 (MBI): From x_k, sample $y_k \sim \pi_\eta^{Y|X}(\cdot, x_k)$ which is a manifold Brownian increment.

Step 2 (RHK): From y_k, sample $x_{k+1} \sim \pi_\eta^{X|Y}(\cdot, y_k) \propto e^{-f(x)}\nu(\eta, x, y_k)$.

end for

as the minimal solution of the following heat equation:

$$\frac{\partial}{\partial t}\nu(t, x, y) = \frac{1}{2}\Delta_y \nu(t, x, y) \qquad \text{with} \qquad \nu(0, x, y) = \delta_x(y).$$

More details can be found in Hsu [2002, Chapter 4]. Unlike the Euclidean case, on Riemannian manifold, the heat kernel does not have a closed-form solution in general. However, some properties of the Euclidean heat kernel is preserved on a Riemannian manifold. One such property is the following: Consider $M = \mathbb{R}^d$ we have $t \log \nu(t, x, y) = t \log \frac{1}{(2\pi t)^{d/2}} - \frac{\|x-y\|^2}{2}$. As $t \to 0$, we get $\lim_{t\to 0} t \log \nu(t, x, y) = -\frac{\|x-y\|^2}{2}$. On a Riemannian manifold, we have the following result.

Fact 5 (Varadhan's asymptotic relation [Hsu, 2002])**.** *For all $x, y \in M$ with $y \notin \mathrm{Cut}\,(x)$, we have*

$$\lim_{t\to 0} t \log \nu(t, x, y) = -\frac{d(x,y)^2}{2} \quad \textit{and} \quad \lim_{t\to 0} t \operatorname{grad}_y \log \nu(t, x, y) = \exp_y^{-1}(x).$$

When evaluation of the heat kernel is required for practical applications, the Varadhan asymptotics aforementioned is used [De Bortoli et al., 2022].

Yet another numerical method for evaluating the heat kernel on manifold is truncation method; see, for example, Corstanje et al. [2024, Section 5.1] and De Bortoli et al. [2022]. In many cases, the heat-kernel has an infinite series expansion. For example, a power series expansion of heat kernel on hypersphere is given in Zhao and Song [2018, Theorem 1], and more examples can be found in Eltzner et al. [2021, Example 1-5]. Similar results are also available for more general manifolds; see, for example, Azangulov et al. [2022] for compact Lie groups and their homogeneous space, and Azangulov et al. [2024] for non-compact symmetric spaces. Hence, a natural approach is to truncate this infinite series at an appropriate level. For example, on $\mathcal{S}^2 \subseteq \mathbb{R}^3$, the heat kernel and its truncation up to the l-th term (denoted as ν_l) can be written respectively as

$$\nu(t, x, y) = \sum_{i=0}^{\infty} e^{-\frac{i(i+1)t}{2}} \frac{2i+1}{4\pi} P_i^0(\langle x, y\rangle_{\mathbb{R}^3}) \text{ and } \nu_l(t, x, y) = \sum_{i=0}^{l} e^{-\frac{i(i+1)t}{2}} \frac{2i+1}{4\pi} P_i^0(\langle x, y\rangle_{\mathbb{R}^3}),$$

where P_i^0 are Legendre polynomials.

3 The Riemannian proximal sampler

We now describe the Riemannian Proximal Sampler, introduced in Algorithm 1. Similar to the Euclidean proximal sampler [Lee et al., 2021], the algorithm has two steps. The first step is sampling from the Manifold Brownian Increment (MBI) oracle. The second step is called the Riemannian Heat-Kernel (RHK) Oracle. Recall that $\nu(\eta, x, y)$ denotes the density of manifold Brownian motion with time η. Define a joint distribution $\pi_\eta(x, y) \propto e^{-f(x)}\nu(\eta, x, y)$. Then, step 2 consists of sampling from the aforementioned distribution. When there is no ambiguity, we omit the step size η and simply write $\pi(x, y) \propto e^{-f(x)}\nu(\eta, x, y)$. Algorithm 1 is an idealized algorithm, in the sense that we assume exact access to MBI and RHK oracles. Following Chen et al. [2022], next we provide an intuitive explanation for the algorithm from a diffusion process perspective.

Step 1: For fixed x, we see that $\pi_\eta^{Y|X}(\cdot, x) \propto \nu(\eta, x, \cdot)$ which is the density of Brownian motion starting from x for time η. From this we see that the first step of the algorithm is running forward manifold heat flow: $dZ_t = dB_t$.

Step 2: We will illustrate that the second step of the algorithm is running the time-reversed process of the forward process. Consider a stochastic process $Z_t : t \geq 0$. When we have observations

of $x_\eta \sim Z_\eta$, we can compute the conditional probability of Z_0 conditioned on endpoint Z_η. We denote $\mu(x_0|x_\eta)$ as the posterior. Bayes Theorem says $\mu(x_0|x_\eta) \propto \mu(x_0)L(x_\eta|x_0)$, where $\mu(x_0)$ is the prior guess and the likelihood L depends on the model. We consider the following model (forward heat flow): $dZ_t = dB_t$ with $Z_0 \sim \pi^X \propto e^{-f(x)}$. Then $\mu(x_0) = \pi^X(x_0)$ and $L(x_\eta|x_0) = \nu(\eta, x_0, x_\eta)$. Thus we get $\mu(x_0|x_\eta) \propto e^{-f(x_0)}\nu(\eta, x_0, x_\eta)$, and we observe that $\mu(x_0|x_\eta)$ is exactly $\pi^{X|Y=x_\eta}(x_0|x_\eta)$. For the forward heat flow $dZ_t = dB_t$ with initialization $Z_0 \sim \pi^X \propto e^{-f(x)}$, there is a well-defined time reversed process $\hat{Z}_t^-$, which satisfies $(Z_0, Z_\eta) \overset{d}{=} (\hat{Z}_\eta^-, \hat{Z}_0^-)$. See Appendix D.2 for more details. Based on this, for the time-reversed process $\hat{Z}_t^-$, the law of $\hat{Z}_\eta^-$ conditioned on $\hat{Z}_0^- = z$ is the same as the posterior $\mu(x|z)$ discussed previously, i.e., $\pi^{X|Y=z}(x) \propto e^{-f(x)}\nu(\eta, x, z)$. Thus we see that the RHK oracle is, from a diffusion perspective, running the time-reversed process.

Implementing Step 1 and Step 2 is non-trivial on Riemannian manifolds. In Section 5 and Appendix A respectively, we discuss two approaches based on heat-kernel truncation and Varadhan's asymptotics. Furthermore, geodesic random walk [Mangoubi and Smith, 2018, Schwarz et al., 2023] is a popular approach to simulate Manifold Brownian Increments (see Appendix B.1), however to the best of our knowledge (in various metrics of interest) is known only under strong assumptions [Cheng et al., 2022, Mangoubi and Smith, 2018].

4 High-accuracy convergence rates

In this section, we provide the convergence rates for the Riemannian Proximal Sampler (Algorithm 1) assuming that the target density satisfies the LSI assumption. Firstly, note that in [Lee et al., 2021] the analysis of Euclidean Proximal Sampler is done assuming the potential function is strongly convex. However, it is known that on a compact manifold, if a function is geodesically convex, then it has to be a constant. Hence assuming the potential f being geodesically convex is not much meaningful. Recently, Cheng et al. [2022] discussed an analog of log-concave distribution on manifolds. Although their setting works for compact manifolds, it requires the Riemannian Hessian of the potential f to be lower bounded by some curvature-related value, which is still restrictive. Hence, we adopt the setting as in Chen et al. [2022], assuming that the target distribution satisfies the LSI.

In Section 4.1, we consider the case where both steps of Algorithm 1 are implemented exactly, and in Section 4.2, we consider the case when MBI and RHK oracles are inexact. Regarding notation, we let $\rho_k^X(x)$, $\rho_k^Y(y)$ denote the law of x and y generated by Algorithm 1 at k-th iteration, assuming exact MBI and exact RHK oracles. When the oracles are inexact, we let $\tilde{\rho}_k^X(x)$, $\tilde{\rho}_k^Y(y)$ to denote the law of x and y generated by Algorithm 1 at k-th iteration.

4.1 Rates with exact oracles

Our first result is as follows, with the proof provided in Appendix D.

Theorem 6. *Let M be a Riemannian manifold without boundary, i.e., $\partial M = \emptyset$. Denote the distribution for the k-th iteration of Algorithm 1 as $x_k \sim \rho_k^X$. Let κ denote the lower bound of Ricci curvature. For any initial distribution ρ_0^X, we have*

1. *Assume π^X satisfies α-LSI.*
   - *For non-negative curvature we have $H_{\pi^X}(\rho_k^X) \le H_{\pi^X}(\rho_0^X)/(1+\eta\alpha)^{2k}, \forall \eta > 0$*
   - *For negative curvature, we have $H_{\pi^X}(\rho_k^X) \le H_{\pi^X}(\rho_0^X)/(1+\eta\frac{\alpha}{2})^{2k}, \forall 0 < \eta \le 1/|\kappa|$.*
2. *Assume π^X satisfies α-PI.*
   - *For non-negative curvature we have $\chi^2_{\pi^X}(\rho_k^X) \le \chi^2_{\pi^X}(\rho_0^X)/(1+\eta\alpha)^{2k}, \forall \eta > 0$*
   - *For negative curvature, we have $\chi^2_{\pi^X}(\rho_k^X) \le \chi^2_{\pi^X}(\rho_0^X)/(1+\eta\frac{\alpha}{2})^{2k}, \forall 0 < \eta \le 1/|\kappa|$.*

Note that the resulting contraction rate depends on the curvature. If the curvature is non-negative, then we can recover the rate in Euclidean space. But in the case of negative curvature, the rate becomes more complicated, and in order to get the contraction rate as in Euclidean space, we need the step size to be bounded above by some curvature-dependent constant.

The above result provides a high-accuracy guarantee for the Riemannian Proximal Sampler in KL-divergence and χ^2 divergence. To see that, consider for example the case when the Ricci curvature is non-negative. Note that to achieve ε accuracy in KL divergence, we need $\frac{H_{\pi^X}(\rho_0^X)}{(1+\eta\alpha)^{2k}} = \varepsilon$. Taking $\log$ on both sides, we get $k = \mathcal{O}(\frac{\log(H_{\pi^X}(\rho_0^X)/\varepsilon)}{\log(1+\eta\alpha)})$. For small step size η, we have $\frac{1}{\log(1+\eta\alpha)} = \mathcal{O}(\frac{1}{\eta\alpha})$. Hence $k = \mathcal{O}(\frac{1}{\eta\alpha}\log\frac{H_{\pi^X}(\rho_0^X)}{\varepsilon}) = \tilde{\mathcal{O}}(\frac{1}{\eta}\log\frac{1}{\varepsilon})$. As η does not depend on ε, we see that we need $\tilde{\mathcal{O}}(\log\frac{1}{\varepsilon})$ number of iterations.

There are several challenges in obtaining the aforementioned result for the Riemannian Proximal Sampler. In Euclidean space, when a probability distribution π^X satisfies α-LSI, its propagation along heat flow $\pi^X * \mathcal{N}(0, tI_d)$ satisfies α_t-LSI, with $\alpha_t = \frac{\alpha}{1+\alpha t}$. This fact is very important and leveraged in Chen et al. [2022] for proving their convergence rates. A quantitative generalization of such a fact for Riemannian manifolds is not immediate and we establish the required results in Appendix H.2, following Collet and Malrieu [2008], under the required Ricci curvature assumptions.

4.2 Rates with inexact oracles

Recall that Algorithm 1 is an idealized algorithm, where we assumed the availability of the MBI and RHK oracles. Note that given $x \in M$, exact MBI oracle generate samples $y \sim \pi_\eta^{Y|X}(\cdot|x)$. And given $y \in M$, exact RHK generate samples $x \sim \pi_\eta^{X|Y}(\cdot|y)$. In practice, exactly implementing these oracles could be computationally expensive or even impossible. For the Euclidean case, we emphasize that, as the heat kernel has an explicit closed form density (which is the Gaussian), prior works, for example, Fan et al. [2023], only consider inexact Restricted Gaussian Oracles and control the propagated error along iterations.

In this section, we derive rates of convergence in the setting where both the MBI and RHK oracles are implemented inexactly. Specifically, we assume we are able to approximately implement the MBI oracle by generating $y \sim \hat{\pi}_\eta^{Y|X}(\cdot|x)$, and approximately implement the RKH oracle by generating $x \sim \hat{\pi}_\eta^{X|Y}(\cdot|y)$, see Assumption 1 below.

Assumption 1. *Denote the output of exact RHK oracle as $\pi_\eta^{X|Y}(\cdot|y)$ and inexact RHK oracle as $\hat{\pi}_\eta^{X|Y}(\cdot|y)$. Similarly, denote the output of exact MBI oracle as $\pi_\eta^{Y|X}(\cdot|x)$ and inexact MBI oracle as $\hat{\pi}_\eta^{Y|X}(\cdot|x)$. Let ζ_{RHK} and ζ_{MBI} be the desired accuracy. We assume that, for inverse step size $\eta^{-1} = \tilde{\mathcal{O}}(\log\frac{1}{\zeta})$, the RHK and MBI oracle implementations can achieve respectively $\|\hat{\pi}_\eta^{X|Y}(\cdot|y) - \pi_\eta^{X|Y}(\cdot|y)\|_{TV} \le \zeta_{\mathsf{RHK}}, \forall y$, and $\|\hat{\pi}_\eta^{Y|X}(\cdot|x) - \pi_\eta^{Y|X}(\cdot|x)\|_{TV} \le \zeta_{\mathsf{MBI}}, \forall x$. We then let $\zeta := \max\{\zeta_{\mathsf{RHK}}, \zeta_{\mathsf{MBI}}\}$.*

The need for assuming the step size satisfies $\eta^{-1} = \tilde{\mathcal{O}}(\log\frac{1}{\zeta})$ for the approximation quality is as follows. Recall from the discussion below Theorem 6 that the complexity of Riemannian Proximal Sampler depends on the step size as $\mathcal{O}(\frac{1}{\eta})$. Thus if η became too small, for example $\eta^{-1} = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$, then the overall complexity would be $\mathrm{Poly}\,(\frac{1}{\varepsilon})$, which is not a high-accuracy guarantee.

We also briefly explain the intuition in assuming total variation distance error bound in oracle quality, and postpone the detailed discussion to Section 5. To guarantee a high quality oracle, we need a high quality approximation of heat kernel. As mentioned previously, a popular method is through truncation of infinite series. Theoretically, the L_2 truncation error can be bounded for compact manifold [Azangulov et al., 2022], which says that the difference between the heat kernel and the approximation of heat kernel are close. This naturally imply an error bound in total variation distance, which motivates us to consider the propagated error in total variation distance.

We first start with a result quantifying the error propagated along iterations, under the availability of inexact oracles. The proof of the following result is provided in Appendix E.

Lemma 7. *Let ρ_k^X denote the law of X through exact oracle implementation of Algorithm 1, and $\tilde{\rho}_k^X$ denote the law of x through inexact oracle implementation of Algorithm 1. Under Assumption 1, we have $\|\rho_k^X(x) - \tilde{\rho}_k^X(x)\|_{TV} \le k(\zeta_{\mathsf{RHK}} + \zeta_{\mathsf{MBI}})$.*

Based on this result, we next obtain the following result analogues to Theorem 6; the proof is provided in Appendix E.

Theorem 8. *Similar to Theorem 6, let M be a Riemannian manifold without boundary. Assume Assumption 1 holds. For any initial distribution ρ_0^X, to reach $\tilde{\mathcal{O}}(\varepsilon)$ total variation distance with oracle accuracy $\zeta = \zeta_{\mathsf{RHK}} = \zeta_{\mathsf{MBI}} = \frac{\varepsilon}{\log^2 \frac{1}{\varepsilon}}$ and step size $\frac{1}{\eta} = \tilde{\mathcal{O}}(\log \frac{1}{\varepsilon})$ (for negative curvature, we additionally require $\eta \leq 1/|\kappa|$),*

1. *if π^X satisfies α-LSI, we need $k = \tilde{\mathcal{O}}(\log^2 \frac{1}{\varepsilon})$ iterations.*
2. *if π^X satisfies α-PI, we need $k = \tilde{\mathcal{O}}\left(\log \frac{1}{\varepsilon} \log \frac{\chi^2_{\pi^X}(\rho_0^X)}{\varepsilon^2}\right) = \tilde{\mathcal{O}}\left(\log^2 \frac{1}{\varepsilon}\right)$ iterations.*

Remarks By Villani [2008, Thm. 6.15], $W_1(\mu, \nu) \leq D\|\mu - \nu\|_{TV}$, where D is the manifold's diameter. For compact manifolds, when D is constant, our TV bound directly implies a W_1 bound.

5 Implementation of inexact oracles via heat kernel trucation

Theorem 8 shows that as long we have sufficient accuracy of MBI and RHK oracles satisfying Assumption 1, we can have a high-accuracy Riemannian sampling algorithm. In this section, we introduce an approximate implementation, based on heat kernel truncation (as introduced in Section 2) and rejection sampling. Numerical simulations for this approach are provided in Appendix B.2.

First note that for rejection sampling method (in general) there are two key ingredients: a proposal distribution and an acceptance rate. Assume we want to generate samples from ρ through rejection sampling. We choose a suitable proposal distribution denoted as μ, and a suitable scaling constant K such that the acceptance rate $K\frac{\rho(x)}{\mu(x)} \leq 1, \forall x$. We generate a random proposal $x \sim \mu$ and $u \in [0, 1]$ being a uniform random number. Then we compute $K\frac{\rho(x)}{\mu(x)}$, and accept x if $u \leq K\frac{\rho(x)}{\mu(x)}$.

We also introduce the following definition of Riemannian Gaussian distribution, as defined next, which will be used as the proposal distribution in rejection sampling. A Riemannian Gaussian distribution centered at x^* with variable t is $\mu(t, x^*, x) \propto \mu_u(t, x^*, x) := \exp\left(-\frac{d(x^*,x)^2}{2t}\right)$, where μ_u denote an unnormalized version of μ and d denotes the geodesic distance. We use this as our proposal distribution to implement rejection sampling, as exact sampling from such a distribution is well-studied for certain specific manifolds; see, for example, Said et al. [2017] for symmetric spaces and Chakraborty and Vemuri [2019] for Stiefel manifolds. Furthermore, this notion of a Riemannian Gaussian distribution is also used in the study of differential privacy on Riemannian manifolds due to their practical feasibility [Reimherr et al., 2021, Jiang et al., 2023]. In section I.2 we provide an explicit algorithm for sampling from the Riemannian Gaussian distribution on the sphere via rejection sampling.

5.1 Implementation of RHK

We first recall the rejection sampling implementation of Restricted Gaussian Oracle (RGO) in the Euclidean setting. Note that, we have $\log \nu_u(\eta, x, y_k) = -\frac{1}{2\eta}\|x - y_k\|^2$, where $\nu_u = \exp(-\frac{1}{2\eta}\|x - y_k\|^2)$ is an unnormalized heat kernel (or the Gaussian density) in Euclidean space. Then we have $\pi_\eta^{X|Y}(\cdot, y_k) \propto e^{-f(x)-\frac{1}{2\eta}\|x-y_k\|^2}$. Then, the RGO is implemented through rejection sampling. Specifically, we can first find the minimizer $x^* \in \arg\min_x f(x) + \frac{1}{2\eta}\|x - y_k\|^2$. Note that the minimizer represents the mode of $\pi_\eta^{X|Y}(\cdot, y_k)$. We can then sample a Gaussian proposal $x_p \sim \mathcal{N}(x^*, tI_d)$ for suitable t centered at the mode x^* and perform rejection sampling. For more details, see, for example, Chewi [2023].

On a Riemannian manifold with ν denoting the heat kernel, to sample from $\pi_\eta^{X|Y}(\cdot, y_k) \propto e^{-f(x)}\nu(\eta, x, y_k)$ through rejection sampling, we need evaluations of $f(x) - \log \nu(\eta, x, y_k)$. But in general, we cannot evaluate the heat kernel exactly, hence we seek for certain heat kernel approximations. Hence, we use the truncated heat kernel ν_l to replace ν, and perform rejection sampling, see Algorithm 2. In the rejection sampling algorithm, as mentioned previously, we use a Riemannian Gaussian distribution as the proposal for rejection sampling. When the minimizer of g is available, we can set x^* to be the minimizer; otherwise, we can simply set $x^* = y_k$. We choose suitable step size η

Algorithm 2 RHK through rejection sampling

Set $x^* = y_k$ and denote $g(x) := f(x) - \log \nu_l(\eta, x, y_k)$.
Set suitable t and constant C_{RHK} s.t. $V_{\mathsf{RHK}}(x) := \frac{\exp(-g(x)+g(x^*)+C_{\mathsf{RHK}})}{\exp(-\frac{1}{2t}d(x,x^*)^2)} \leq 1, \forall x \in M$
for $i = 0, 1, 2, ...$ **do**
  Generate proposal $x \sim \mu(t, x^*, \cdot)$.
  Generate u uniformly on $[0, 1]$.
  Return x if $u \leq V_{\mathsf{RHK}}(x)$
end for

Algorithm 3 MBI through rejection sampling

Set suitable t and C_{MBI} so that $V_{\mathsf{MBI}}(y) := \frac{\exp(\log \nu_l(\eta,x,y) - \log \nu_l(\eta,x,x) + C_{\mathsf{MBI}})}{\exp(-\frac{d(x,y)^2}{2t})} \leq 1, \forall y \in M$
for $i = 0, 1, 2, ...$ **do**
  Generate proposal $y \sim \mu(t, x, \cdot)$.
  Generate u uniformly on $[0, 1]$.
  Return y if $u \leq V_{\mathsf{MBI}}(y)$
end for

and t that depends on η s.t. $g(x) - g(x^*) + C_{\mathsf{RHK}} \geq \frac{1}{2t}d(x, x^*)^2$. Such an inequality can guarantee that the acceptance rate (with Riemannian Gaussian distribution $\mu(t, x^*, x)$ as proposal) would not exceed one, i.e., $V_{\mathsf{RHK}}(x) \leq 1, \forall x$. Then we see that the output of rejection sampling would follow $\hat{\pi}_\eta^{X|Y}(x|y_k) \propto \exp(f(x) - \log \nu_l(\eta, x, y_k))$. Similarly, to implement the MBI oracle, we also use rejection sampling to get a high-accuracy approximation. Specifically, Algorithm 3 generates inexact Brownian motion starting from x with time η.

5.2 Verification of Assumption 1

We now show that Assumption 1 is satisfied for the aforementioned inexact implementation of the Riemannian Proximal Sampler. To do so, we specifically consider the case when the manifold M is compact and is a homogeneous space. Recall that ν_l denote the truncated heat kernel with truncation level l. Roughly speaking, a homogeneous space is a manifold that has certain symmetry, including Stiefel manifold, Grassmann manifold, hypersphere, and manifold of positive definite matrices.

Proposition 9. *Let M be a compact manifold. Assume further that M is a homogeneous space. With truncation implementation of inexact oracles, in order for Assumption 1 to be satisfied with $\zeta = \frac{\varepsilon}{\log^2 \frac{1}{\varepsilon}}$, we need truncation level l to be of order polylog$(1/\varepsilon)$.*

Sketch of proof: We briefly mention the idea of proof. Azangulov et al. [2022, Proposition 21] provided an L_2 bound on the truncation error, and by Jensen's inequality we get an L_1 bound as desired. With truncation level l to be of order Poly $(\log \frac{1}{\varepsilon})$, we can achieve $\int_M |\nu(\eta, x, y) - \nu_l(\eta, x, y)| dV_g(x) = \tilde{\mathcal{O}}(\zeta)$. See Proposition 19 and Proposition 22 for a complete proof.

Remark. Proposition 9 concerns Algorithms 2 and 3, which is one way to implement the RHK and MBI oracles. Rejection sampling is a part of the implementation in Algorithm 2 and 3, which are both based on truncated heat kernels. The cost of rejection sampling comes in the number of steps of the **for** loop in both the algorithms. Intuitively, one can expect that if we have a highly-accurate evaluation of the heat kernel, the cost of rejection sampling should be the same as that of rejection sampling with the exact heat kernel.

Remark. On the Euclidean space, for which the exact heat kernel is known, the cost of rejection sampling can be proved to be $\mathcal{O}(1)$ Chen et al. [2022]. Hypothetically, even if we have the exact heat kernel on a Riemannian manifold, the cost for rejection sampling is actually unknown for general Riemannian manifolds. For the case of sphere, we provide an end-to-end result (including the cost of rejection sampling) in Corollary 10. In proving this result, we first showed that when the acceptance rate V in rejection sampling would possibly exceed 1 in some unimportant regions, Assumption 1 still holds, via explicit computations (see Appendix I.1). Then, we show that the cost of rejection sampling (even with the inexact heat-kernel based on truncation level as stated in Proposition 9), is $\mathcal{O}(1)$ similar to the Euclidean case.

When M is not a homogeneous space, to the best of our knowledge, it is unknown how to implement the truncation method. Exploring this direction to further extend the above result is an interesting direction for future work.

5.3 A concrete example on hyperspheres

We provide a more specific computational complexity result that consider the dimension dependency as well as cost for rejection sampling; the proof is provided in Appendix I.1.

Corollary 10. *Let $M = \mathcal{S}^d$, and let π^X satisfies α-LSI with the potential function f additionally being L_1-Lipschitz on M. Assume without the loss of generality that $L_1 \geq \sqrt{d}$. Consider heat kernel truncation implementation (i.e., Algorithm 2 and 3), without minimization (i.e., start rejection sampling from y_k directly), and with step size $\eta = \frac{1}{L_1^2 d \log \frac{L_1^2 d \log^2 \frac{1}{\varepsilon}}{\varepsilon}}$ and truncation level $l = \mathcal{O}(d^2 \operatorname{Poly}(\log \frac{1}{\varepsilon}))$. To get an ϵ-accurate sample in TV distance, the iteration complexity, is $k = \tilde{\mathcal{O}}(\frac{L_1^2 d}{\alpha} \log^2 \frac{1}{\varepsilon})$, where we use $\tilde{\mathcal{O}}$ to keep only the leading factors.*

6 Additional results

- In Setion A, we design another practical implementations of the oracles based on Varadhan's asymptotics. While showing that this implementation satisfies Assumption 1 is left as future work, we show their connection to entropy-regularized proximal point methods on Wasserstein spaces (see Theorem 12).
- We evaluate the empirical performance of both implementations through simulation studies presented in Appendix B.

7 Concluding remarks

We introduced the *Riemannian Proximal Sampler* for sampling from densities on Riemannian manifolds. By leveraging the Manifold Brownian Increments (MBI) and the Riemannian Heat-kernel (RHK) oracles, we established high-accuracy sampling guarantees, demonstrating a logarithmic dependence on the inverse accuracy parameter (i.e., polylog$(1/\varepsilon)$) in the Kullback-Leibler divergence (for exact oracles) and total variation metric (for inexact oracles). Additionally, we proposed practical implementations of these oracles using heat-kernel truncation and Varadhan's asymptotics, providing a connection between our sampling method and the Riemannian Proximal Point Method.

Future works include: (i) characterizing the precise dependency on other problem parameters apart from ε, (ii) improving oracle approximations for enhanced computational efficiency and (iii) extending these techniques to broader classes of manifolds (and other metric-measure spaces).

Acknowledgments and Disclosure of Funding

Krishnakumar Balasubramanian was supported in part by NSF grant DMS-2413426. Shiqian Ma was supported in part by ONR grant N00014-24-1-2705, NSF grants CCF-2311275 and ECCS-2326591.

References

Jason M Altschuler and Sinho Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. *Journal of the ACM*, 71(3):1–55, 2024.

Christophe Andrieu, Anthony Lee, Sam Power, and Andi Q Wang. Explicit convergence bounds for Metropolis Markov chains: Isoperimetry, spectral gaps and profiles. *The Annals of Applied Probability*, 34(4):4022–4071, 2024.

Alexis Arnaudon, Alessandro Barp, and So Takao. Irreversible Langevin MCMC on lie groups. In *Geometric Science of Information: 4th International Conference, GSI 2019, Toulouse, France, August 27–29, 2019, Proceedings 4*, pages 171–179. Springer, 2019.

Iskander Azangulov, Andrei Smolensky, Alexander Terenin, and Viacheslav Borovitskiy. Stationary Kernels and Gaussian Processes on Lie Groups and their Homogeneous Spaces I: the compact case. *arXiv e-prints*, pages arXiv–2208, 2022.

Iskander Azangulov, Andrei Smolensky, Alexander Terenin, and Viacheslav Borovitskiy. Stationary Kernels and Gaussian Processes on Lie Groups and their Homogeneous Spaces II: non-compact symmetric spaces. *Journal of Machine Learning Research*, 25(281):1–51, 2024.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

Karthik Bharath, Alexander Lewis, Akash Sharma, and Michael V Tretyakov. Sampling and estimation on manifolds using the Langevin diffusion. *arXiv preprint arXiv:2312.14882*, 2023.

Clément Bonet, Paul Berg, Nicolas Courty, François Septier, Lucas Drumetz, and Minh Tan Pham. Spherical Sliced-Wasserstein. In *The Eleventh International Conference on Learning Representations*, 2023.

Simon Byrne and Mark Girolami. Geodesic Monte Carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40(4):825–845, 2013.

Djalil Chafaï. Entropies, convexity, and functional inequalities: On phi-entropies and phi-sobolev inequalities. *Journal of Mathematics of Kyoto University*, 44(2):325–363, 2004.

Rudrasis Chakraborty and Baba C Vemuri. Statistics on the Stiefel manifold: Theory and applications. *The Annals of Statistics*, 47, 2019.

August Y Chen and Karthik Sridharan. Optimization, isoperimetric inequalities, and sampling via lyapunov potentials. *arXiv preprint arXiv:2410.02979*, 2024.

Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, pages 2984–3014. PMLR, 2022.

Yuansi Chen and Khashayar Gatmiry. When does Metropolized Hamiltonian Monte Carlo provably outperform Metropolis-adjusted Langevin algorithm? *arXiv preprint arXiv:2304.04724*, 2023.

Yuansi Chen, Raaz Dwivedi, Martin J Wainwright, and Bin Yu. Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients. *Journal of Machine Learning Research*, 21(92):1–72, 2020.

Xiang Cheng, Jingzhao Zhang, and Suvrit Sra. Efficient sampling on Riemannian manifolds via Langevin MCMC. *Advances in Neural Information Processing Systems*, 35:5995–6006, 2022.

Sinho Chewi. Log-concave sampling. *Book draft available at* `https://chewisinho.github.io`, 2023.

Sinho Chewi and Austin J Stromme. The ballistic limit of the log-sobolev constant equals the Polyak–łojasiewicz constant. *arXiv preprint arXiv:2411.11415*, 2024.

Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. Optimal dimension dependence of the Metropolis-adjusted Langevin algorithm. In *Conference on Learning Theory*, pages 1260–1300. PMLR, 2021.

Jean-François Collet and Florent Malrieu. Logarithmic Sobolev inequalities for inhomogeneous Markov semigroups. *ESAIM: Probability and Statistics*, 12:492–504, 2008.

Marc Corstanje, Frank van der Meulen, Moritz Schauer, and Stefan Sommer. Simulating conditioned diffusions on manifolds. *arXiv preprint arXiv:2403.05409*, 2024.

Christopher Criscitiello and Nicolas Boumal. An accelerated first-order method for non-convex optimization on manifolds. *Foundations of Computational Mathematics*, 23(4):1433–1509, 2023.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. *Advances in Neural Information Processing Systems*, 35:2406–2422, 2022.

Paromita Dubey and Hans-Georg Müller. Fréchet analysis of variance for random objects. *Biometrika*, 106(4):803–821, 2019.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast. *Journal of Machine Learning Research*, 20(183):1–42, 2019.

Benjamin Eltzner, Pernille Hansen, Stephan F Huckemann, and Stefan Sommer. Diffusion means in geometric spaces. *arXiv preprint arXiv:2105.12061*, 2021.

Jiaojiao Fan, Bo Yuan, and Yongxin Chen. Improved dimension dependence of a proximal algorithm for sampling. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1473–1521. PMLR, 2023.

Maurice Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l'institut Henri Poincaré*, volume 10, pages 215–310, 1948.

Khashayar Gatmiry and Santosh S Vempala. Convergence of the Riemannian Langevin Algorithm. *arXiv preprint arXiv:2204.10818*, 2022.

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.

Yun Gong, Niao He, and Zebang Shen. Poincare inequality for local log-polyak-\l ojasiewicz measures: Non-asymptotic analysis in low-temperature regime. *arXiv preprint arXiv:2501.00429*, 2024.

Navin Goyal and Abhishek Shetty. Sampling and optimization on convex sets in Riemannian manifolds of non-negative curvature. In *Conference on Learning Theory*, pages 1519–1561. PMLR, 2019.

Ye He, Alireza Mousavi-Hosseini, Krishnakumar Balasubramanian, and Murat A Erdogdu. A Separation in Heavy-Tailed Sampling: Gaussian vs. Stable Oracles for Proximal Samplers. *arXiv preprint arXiv:2405.16736*, 2024.

Elton P Hsu. Logarithmic Sobolev inequalities on path spaces over Riemannian manifolds. *Communications in mathematical physics*, 189(1):9–16, 1997.

Elton P Hsu. *Stochastic analysis on manifolds*. Number 38 in Graduate Studies in Mathematics,. American Mathematical Soc., 2002.

Chin-Wei Huang, Milad Aghajohari, Joey Bose, Prakash Panangaden, and Aaron C Courville. Riemannian diffusion models. *Advances in Neural Information Processing Systems*, 35:2750–2761, 2022.

Yangdi Jiang, Xiaotian Chang, Yi Liu, Lei Ding, Linglong Kong, and Bei Jiang. Gaussian differential privacy on Riemannian manifolds. *Advances in Neural Information Processing Systems*, 36: 14665–14684, 2023.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Ravi Kannan, László Lovász, and Miklós Simonovits. Random walks and an o*(n5) volume algorithm for convex bodies. *Random Structures & Algorithms*, 11(1):1–50, 1997.

Ravi Kannan, László Lovász, and Ravi Montenegro. Blocking conductance and mixing in random walks. *Combinatorics, Probability and Computing*, 15(4):541–570, 2006.

Lingkai Kong and Molei Tao. Convergence of kinetic Langevin Monte Carlo on lie groups. *arXiv preprint arXiv:2403.12012*, 2024.

Yunbum Kook and Santosh S Vempala. Sampling and integration of logconcave functions by algorithmic diffusion. *arXiv preprint arXiv:2411.13462*, 2024.

Yunbum Kook and Matthew S Zhang. Rényi-infinity constrained sampling with d^3 membership queries. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 5278–5306. SIAM, 2025.

Yunbum Kook, Yin-Tat Lee, Ruoqi Shen, and Santosh Vempala. Sampling with Riemannian Hamiltonian Monte Carlo in a constrained space. *Advances in Neural Information Processing Systems*, 35:31684–31696, 2022.

Yunbum Kook, Santosh S Vempala, and Matthew S Zhang. In-and-Out: Algorithmic Diffusion for Sampling Convex Bodies. *arXiv preprint arXiv:2405.01425*, 2024.

John M Lee. *Introduction to Riemannian manifolds*, volume 2. Springer, 2018.

Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Logsmooth gradient concentration and tighter runtimes for Metropolized Hamiltonian Monte Carlo. In *Conference on learning theory*, pages 2565–2597. PMLR, 2020.

Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured logconcave sampling with a restricted Gaussian oracle. In *Conference on Learning Theory*, pages 2993–3050. PMLR, 2021.

Benedict Leimkuhler and Charles Matthews. Efficient molecular dynamics using geodesic integration and solvent–solute splitting. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 472(2189):20160138, 2016.

Mufan Li and Murat A Erdogdu. Riemannian Langevin algorithm for solving semidefinite programs. *Bernoulli*, 29(4):3093–3113, 2023.

Han Cheng Lie, Daniel Rudolf, Björn Sprungk, and Timothy J Sullivan. Dimension-independent Markov chain Monte Carlo on the sphere. *Scandinavian Journal of Statistics*, 50(4):1818–1858, 2023.

Chang Liu and Jun Zhu. Riemannian Stein variational gradient descent for Bayesian inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Chang Liu, Jun Zhu, and Yang Song. Stochastic gradient geodesic MCMC methods. *Advances in neural information processing systems*, 29, 2016.

László Lovász. Hit-and-run mixes fast. *Mathematical programming*, 86:443–461, 1999.

Oren Mangoubi and Aaron Smith. Rapid mixing of geodesic walks on manifolds with positive curvature. *The Annals of Applied Probability*, 28(4):2501–2543, 2018.

Michelle Muniz, Matthias Ehrhardt, Michael Günther, and Renate Winkler. Higher strong order methods for linear Itô SDEs on matrix Lie groups. *BIT Numerical Mathematics*, 62(4):1095–1119, 2022.

Tomohiro Nishiyama and Igal Sason. On relations between the relative entropy and χ 2-divergence, generalizations and applications. *Entropy*, 22(5):563, 2020.

Maxence Noble, Valentin De Bortoli, and Alain Durmus. Unbiased constrained sampling with self-concordant barrier Hamiltonian Monte Carlo. *Advances in Neural Information Processing Systems*, 36:32672–32719, 2023.

Adam Nowak. Personal Communication, 2025.

Adam Nowak, Peter Sjögren, and Tomasz Z Szarek. Sharp estimates of the spherical heat kernel. *Journal de Mathématiques Pures et Appliquées*, 129:23–33, 2019.

Sam Patterson and Yee Whye Teh. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. *Advances in neural information processing systems*, 26, 2013.

Gabriel Peyré. Entropic approximation of Wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.

Marc J Piggott and Victor Solo. Geometric Euler–Maruyama Schemes for Stochastic Differential Equations in SO (n) and SE (n). *SIAM Journal on Numerical Analysis*, 54(4):2490–2516, 2016.

Matthew Reimherr, Karthik Bharath, and Carlos Soto. Differential privacy over Riemannian manifolds. *Advances in Neural Information Processing Systems*, 34:12292–12303, 2021.

Salem Said, Hatem Hajri, Lionel Bombrun, and Baba C Vemuri. Gaussian distributions on Riemannian symmetric spaces: statistical learning with structured covariance matrices. *IEEE Transactions on Information Theory*, 64(2):752–772, 2017.

Simon Schwarz, Michael Herrmann, Anja Sturm, and Max Wardetzky. Efficient random walks on Riemannian manifolds. *Foundations of Computational Mathematics*, pages 1–17, 2023.

Vishwak Srinivasan, Andre Wibisono, and Ashia Wilson. Fast sampling from constrained spaces using the Metropolis-adjusted Mirror Langevin algorithm. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4593–4635. PMLR, 2024a.

Vishwak Srinivasan, Andre Wibisono, and Ashia Wilson. High-accuracy sampling from constrained spaces with the Metropolis-adjusted Preconditioned Langevin Algorithm. *arXiv preprint arXiv:2412.18701*, 2024b.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2003.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2008.

Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, pages 2093–3027. PMLR, 2018.

Keru Wu, Scott Schmidler, and Yuansi Chen. Minimax mixing time of the Metropolis-adjusted Langevin algorithm for log-concave sampling. *Journal of Machine Learning Research*, 23(270): 1–63, 2022.

Xiangjin Xu. Heat kernel gaussian bounds on manifolds i: manifolds with non-negative ricci curvature. *arXiv preprint arXiv:1912.12758*, 2019.

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.

Chenchao Zhao and Jun S Song. Exact heat kernel on a hypersphere and its applications in kernel SVM. *Frontiers in Applied Mathematics and Statistics*, 4:1, 2018.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, see the end of Section 1 for a summary of our main contributions, as well as references on the theorem number.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See section 7 for a discussion on future works.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We did state the assumptions in each theorem, and the proof can be found in appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The information needed to reproduce the toy experiments are provided.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Codes were provided in supplementary material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The details that are necessary to understand the results are provided in this paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Our work is theoretical and includes some toy examples. The plots are average over 1000 number of trails.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The toy examples are run on a personal laptop using Matlab, and only CPU (AMD Ryzen 7 PRO 5850U).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have read the NeurIPS Code of Ethics and followed it in the paper.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our work is theoretical.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work is theoretical and poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our work is theoretical.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our work is theoretical and does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our work is theoretical, and does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Implementation via Varadhan's asymptotics and connection to entropy-regularized JKO scheme

In this section, we consider yet another approximation scheme for implementing Algorithm 1, motivated by its connection with the proximal point method in optimization, where the latter is in the sense of optimization over Wasserstein space[1] [Jordan et al., 1998, Wibisono, 2018, Chen et al., 2022]. Note that the proximal point method is usually called as the JKO scheme after the authors of Jordan et al. [1998].

Specifically, we consider approximating the heat kernel through Varadhan's asymptotics. Let $\hat{\nu}(\eta, x, y) \propto_y \exp(-\frac{d(x,y)^2}{2\eta}) =: \hat{\nu}_u(\eta, x, y)$ be an inexact evaluation of heat kernel. According to Varadhan's asymptotics, $\lim_{\eta\to 0} \hat{\nu}(\eta, x, y) = \nu(\eta, x, y)$. Hence when η is small, $\hat{\nu}$ is a good approximation of the heat kernel. Note that $\hat{\nu}(\eta, x, \cdot)$ in Varadhan's asymptotic is exactly the Riemannian Gaussian distribution $\mu(\eta, x, \cdot)$. Denote $\tilde{\pi}(x, y) = \exp(-f(x) - \frac{d(x,y)^2}{2\eta})$. With inexact MBI implemented through Riemannian Gaussian distribution and inexact RHK implemented through rejection sampling (Algorithm 2) to generate $\tilde{\pi}^{X|Y}(x|y) \propto \exp(-f(x) - \frac{d(x,y)^2}{2\eta})$, we obtain Algorithm 4.

For the case when $M = \mathcal{S}^d$, we prove in Appendix I that to sample from $\tilde{\pi}^{X|Y}(x|y)$ through rejection sampling, with suitable parameters, the cost is $\mathcal{O}(1)$ in both dimension d and step size η. Obtaining similar results for more general manifolds seems non-trivial. Numerical simulations for this approach are provided in Appendix B.3. Verifying Assumption 1 for this implementation is open.

Algorithm 4 Inexact manifold proximal sampler with Varadhan's asymptotics

for $k = 0, 1, 2, ...$ **do**

From x_k, sample $y_k \sim \tilde{\pi}^{Y|X}(\cdot, x_k)$ which is a Riemannian Gaussian distribution.

From y_k, sample $x_{k+1} \sim \tilde{\pi}^{X|Y}(\cdot, y_k) \propto e^{-f(x)-\frac{d(x,y_k^2)}{2\eta}}$ using Algorithm 2.

end for

A.1 RHK as a proximal operator on Wasserstein space

We first show that the inexact RHK output in Algorithm 4 can be viewed as a proximal operator on Wasserstein space, generalizing the Euclidean result in Chen et al. [2020] to the Riemannian setting. Recall that with a function f and d being a distance function, $\text{prox}_{\eta f}(y) = \arg\min_x f(x) + \frac{1}{2\eta} d(x, y)^2$. The (approximated) joint distribution is $\tilde{\pi}(x, y) = \exp(-f(x) - \frac{d(x,y)^2}{2\eta})$. By direct computation we have the following Lemma (proved in Appendix G).

Lemma 11. *We have that*

$$\tilde{\pi}^{X|Y=y} = \arg\min_{\rho\in\mathcal{P}_2(M)} H_{\tilde{\pi}^X}(\rho) + \frac{1}{2\eta} W_2^2(\rho, \delta_y) = \text{prox}_{\eta H_{\tilde{\pi}^X}}(\delta_y),$$

which shows that the ineact RHK implementation is a proximal operator, i.e., $\tilde{\pi}^{X|Y=y} = \text{prox}_{\eta H_{\tilde{\pi}^X}}(\delta_y)$.

A.2 Connection to entropy-regularized JKO scheme

Observe that in Algorithm 4, the Riemannian Gaussian involves distance square, which naturally relates to Wasserstein distance. Now, recall that for a function F in the Wasserstein space, its Wasserstein gradient flow can be approximated through the following discrete time JKO scheme [Jordan et al., 1998]:

$$\rho_{k+1} = \arg\min_{\rho\in\mathcal{P}(\mathbb{R}^d)} F(\rho) + \frac{1}{2\eta} W_2^2(\rho, \rho_k).$$

[1] If M is a smooth compact Riemannian manifold then the Wasserstein space $\mathcal{P}_2(M)$ is the space of Borel probability measures on M, equipped with the Wasserstein metric W_2. We refer the reader to Villani [2003] for background on Wasserstein spaces.

It was proved that as $\eta \to 0$, the discrete time sequence $\{\rho_k\}$ converge to the Wasserstein gradient flow of F. Later, Peyré [2015] proposed an approximation scheme through entropic smoothing of Wasserstein distance:

$$\rho_{k+1} = \operatorname*{arg\,min}_{\rho \in \mathcal{P}(\mathbb{R}^d)} F(\rho) + \frac{1}{2\eta} W_{2,\varepsilon}^2(\rho, \rho_k),$$

where $W_{2,\varepsilon}$ is the entropy-regularized 2-Wasserstein distance defined by (here H is the negative entropy)

$$W_{2,t}^2(\rho_1, \rho_2) = \inf_{\gamma \in \mathcal{C}(\rho_1, \rho_2)} \int d(x,y)^2 d\gamma(x,y) + tH(\gamma).$$

In Euclidean space, Chen et al. [2022] showed that the proximal sampler can be viewed as an entropy-regularized JKO scheme. We extend such an interpretation to Riemannian manifolds. Specifically, we show that Algorithm 4 which is an approximation of the exact proximal sampler (Algorithm 1), can be viewed as an entropy-regularized JKO as stated in Theorem 12 (proved in Appendix G). Note that on a Riemannian manifold the negative entropy is $H(\gamma) := \int_{M \times M} \gamma \log(\gamma) dV_g(x) dV_g(y)$.

Theorem 12. *Recall that $\pi^X \propto e^{-f}$. Let x_k, y_k, x_{k+1} be generated by Algorithm 4. Let $\tilde{\rho}_k^X$, $\tilde{\rho}_k^Y$ and $\tilde{\rho}_{k+1}^X$ be the distribution of x_k, y_k, x_{k+1}, respectively. Then*

$$\tilde{\rho}_k^Y = \operatorname*{arg\,min}_{\chi \in \mathcal{P}_2(M)} \frac{1}{2\eta} W_{2,2\eta}^2(\tilde{\rho}_k^X, \chi) \quad \textit{and} \quad \tilde{\rho}_{k+1}^X = \operatorname*{arg\,min}_{\chi \in \mathcal{P}_2(M)} \int f d\chi + \frac{1}{2\eta} W_{2,2\eta}^2(\tilde{\rho}_k^Y, \chi).$$

B Simulation results

B.1 Brownian motion approximation via geodesic random walk

In our experiments, to compare against the Riemannian Langevin Monte Carlo Algorithm, we used the geodesic random walk algorithm to simulate the MBI oracle following Cheng et al. [2022], De Bortoli et al. [2022], Schwarz et al. [2023]; see Algorithm 5. More efficient implementation is a topic of great interest in the literature; see, for example, [Schwarz et al., 2023].

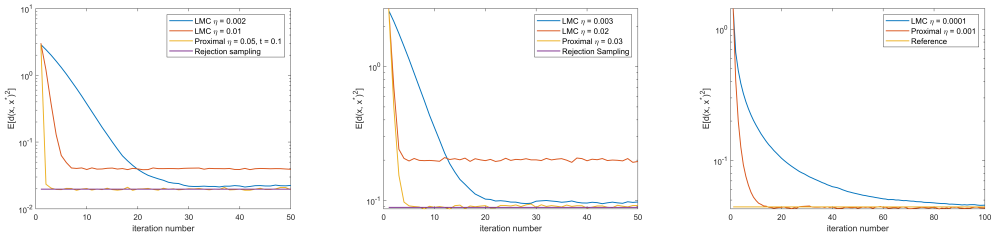
Algorithm 5 Approximation of manifold Brownian motion using geodesic random walk

Input $x \in M, t > 0$.
Sample ξ being a Euclidean Brownian increment with time t in the tangent space $T_x M$.
Output $y = \exp_x(\xi)$.

While it is well-known that geodesic random walks converge asymptotically to the Brownian motion on the manifold, non-asymptotic rates of convergence in various metrics of interest is largely unknown. A basic non-asymptotic error bound for geodesic random walk is available in Wasserstein distance (see Cheng et al. [2022, Lemma 7]). Mixing time results are provided in Mangoubi and Smith [2018]. However, such a result is not immediately applicable to establish high-accuracy guarantees for the Riemannian proximal sampler, when the MBI oracle is implemented via geodesic random walk. An important and interesting future work is establishing rates of convergence for geodesic random walk in various metrics of interest so that those results could be leveraged to obtain high-accuracy guarantees for the Riemannian proximal sampler.

B.2 Numerical experiments for Algorithms 2 and 3: von Mises-Fisher distribution on hyperspheres

In this experiment, we test the performance of Algorithms 2 and 3 for sampling from the von Mises-Fisher distribution on hyperspheres and compare it with the Riemannian LMC method. In this case, we have $f(x) = -\kappa \mu^T x$. Note that this $f(x)$ has a unique minimizer on $\mathcal{S}^d$. This implies that LSI is satisfied, see [Li and Erdogdu, 2023, Theorem 3.4]. We demonstrate the performance of our Algorithm on $\mathcal{S}^2 \subseteq \mathbb{R}^3$ with $\mu = (10, 0.1, 2)^T$ and $\kappa = 10$, and on $\mathcal{S}^5$ with $\mu = (5, 0.1, 2, 1, 1, 1)^T$ and $\kappa = 10$. For the purpose of numerical demonstration, we sample the Riemannian Gaussian distribution through rejection sampling.



(a) $\mathcal{S}^2$, $f(x) = \exp(\kappa\mu^T x)$ (b) $\mathcal{S}^5$, $f(x) = \exp(\kappa\mu^T x)$ (c) P_3, $f(X) = \exp(\frac{-d(X,I)^4}{2\sigma^2})$,

Figure 1: Frechét variance (i.e., $\mathbb{E}[d(x, x^*)^2]$ versus number of iterations. Left and Middle figure correspond to the implementation via Algorithm 2 and 3. Right figure corresponds to implementation via Algorithm 4.

To evaluate the performance, we estimate $\mathbb{E}[d(x, x^*)^2]$, where x^* is the minimizer of f, representing the mode of the distribution, and plot it as a function of iterations. Note that the quantity $\mathbb{E}[d(x, x^*)^2]$ is referred to as Fréchet variance Fréchet [1948], Dubey and Müller [2019]. For this, we generate 1000 samples (by generating samples independently via different runs) and compute $\frac{1}{1000}\sum_{i=1}^{1000} d(x_i, x^*)^2$. We use rejection sampling to generate unbiased samples and get an estimation of the true value. Due to the biased nature of the Riemannian LMC method, to achieve a high accuracy we need a small step size. Contrary to the Riemanian LMC method, the proximal sampler is unbiased, and it can achieve an accuracy while using a large step size and a smaller number of iterations; see Figures 1-(a) and (b). For both algorithms, we use uniform distribution on the hypersphere as initialization.

B.3 Numerical experiments for Algorithm 4: manifold of positive definite matrices

In this subsection we illustrate the performance of Algorithm 4 for sampling on the manifold of positive definite matrices. Let $P_m = \{X \in GL(m) : X^T = X \quad \text{and} \quad y^T X y > 0, \forall y \in \mathbb{R}^m\}$ be the set of symmetric positive definite matrices. According to Bharath et al. [2023, Section 6.2], we can choose $g(U, V) = \text{tr}\,(X^{-1}UX^{-1}V)$ and make (P_m, g) a Riemannian manifold. It is a non-compact manifold with non-positive sectional curvature, geodesically complete and is a homogeneous space of general linear group $GL(m)$. Additional details are provided in Appendix C.3.

We test the performance of Algorithm 4 when the potential function $f(X) = \frac{1}{2\sigma^2} d(X, I_m)^4$, $m = 3, \sigma = 0.03$, following Bharath et al. [2023]. Note that f is not gradient Lipschitz. In the Figure 1-c, we estimate $\mathbb{E}[d(x, x^*)^2]$ and plot it as function of iterations, where $x^* = I_3$ is the minimizer of f, representing the mode of the distribution. For a baseline comparison, we run Riemannian Langevin Monte Carlo for 200 iterations with decreasing step size to get a reference value of $\mathbb{E}[d(x, x^*)^2]$, which serves as the true $\mathbb{E}[d(x, x^*)^2]$. Similar to the previous experiment, we generate 1000 samples from independent run, and compute $\frac{1}{1000}\sum_{i=1}^{1000} d(x_i, x^*)^2$ for each method. For the Riemannian Langevin Monte Carlo method, we find that if we set step size to 0.001 instead of 0.0001, after a few iterations the algorithm diverges (potentially due to lack of gradient Lipschitz condition). But for the proximal sampler (which is an unbiased algorithm), even with a large step size as illustrated in the plots, the approximation scheme still works well and can achieve a higher accuracy than the Riemannian LMC algorithm. For both algorithms we initialize the algorithm at $X_0 = 2I_3$. Note that if we use a random initialization, the Riemannian LMC algorithm (prior work) might diverge (potentially due to lack of gradient Lipschitz condition in our potential).

B.4 Numerical experiments for Algorithm 4: cost of rejection sampling for higer dimensional case

We also demonstrate that the cost for rejection sampling is not exploded by dimension. Following the same setup as Appendix B.2, we consider sampling from a von Mises-Fisher distribution on $\mathcal{S}^{100}$ with $\mu = (10, 0.1, 2, 1, 1, ..., 1)$ and $\kappa = 10$. We test the performance of Algorithm 4 in terms of rejection sampling cost. We choose $\eta = 0.0001$, and compute the average number of iterations executed by rejection sampling. With a total number of rejection sampling oracle being 1×10^6, the average rejection sampling cost is found to be 2.661.

C Additional preliminaries

C.1 Divergence

We will briefly discuss divergence for the manifold setting. More details can be found in Lee [2018]. Recall that in Euclidean space, for a vector field $F = (F_1, ..., F_n)$ in $\mathbb{R}^n$, divergence of F is defined as $\nabla \cdot F = \sum_{i=1}^n \frac{\partial F_i}{\partial x_i}$. It has a natural generalization to the manifold setting using interior multiplication and exterior derivative.

The Riemannian divergence is defined as the function such that $d(i_X(dV_g)) = (\text{div } X)dV_g$, where X is any smooth vector field on M, i denotes interior multiplication and d denotes exterior derivative. See for example Lee [2018, Appendix B] for more details. On a Riemannian manifold, recall the volumn form is $dV_g = \sqrt{\det(g_{ij})}dx^1 \wedge ... \wedge dx^n$. Let $Y = \sum_{i=1}^n Y^i \frac{\partial}{\partial x_i}$. We can compute the interior multiplication as

$$\begin{aligned} i_Y(dV_g) &= \sqrt{\det(g_{ij})} \sum_{j=1}^n ((-1)^{j+1} dx^j(Y)) dx^1 \wedge ... \wedge d\hat{x}^j \wedge ... \wedge dx^n \\ &= \sqrt{\det(g_{ij})} \sum_{j=1}^n ((-1)^{j+1} Y^j) dx^1 \wedge ... \wedge d\hat{x}^j \wedge ... \wedge dx^n. \end{aligned}$$

We can then compute its exterior derivative as

$$\begin{aligned} d(i_Y(dV_g)) &= \sum_{j=1}^n ((-1)^{j+1} \frac{\partial (Y^j \sqrt{\det(g_{ij})})}{\partial x_j} dx^j) dx^1 \wedge ... \wedge d\hat{x}^j \wedge ... \wedge dx^n \\ &= \frac{1}{\sqrt{\det(g_{ij})}} \sum_{j=1}^n \frac{\partial (Y^j \sqrt{\det(g_{ij})})}{\partial x_j} \sqrt{\det(g_{ij})} dx^1 \wedge ... \wedge dx^n \\ &= \frac{1}{\sqrt{\det(g_{ij})}} \sum_{j=1}^n \frac{\partial (Y^j \sqrt{\det(g_{ij})})}{\partial x_j} dV_g. \end{aligned}$$

Hence we get $\text{div}\,(Y) = \frac{1}{\sqrt{\det(g_{ij})}} \sum_{j=1}^n \frac{\partial (Y^j \sqrt{\det(g_{ij})})}{\partial x_j}$. In Euclidean space, this reduces to $\text{div}\,(Y) = \sum_{j=1}^n \frac{\partial Y^j}{\partial x_j}$.

For $u \in C^\infty(M)$ and $X \in \mathfrak{X}(M)$, the divergence operator satisfies the following product rule

$$\text{div}\,(uX) = u\,\text{div}\,(X) + \langle \text{grad } u, X \rangle_g.$$

Furthermore, we have the "integration by parts" formula (with $\tilde{g}$ denote the induced Riemannian metric on ∂M)

$$\int_M \langle \text{grad } u, X \rangle_g dV_g = \int_{\partial M} u \langle X, N \rangle_g dV_{\tilde{g}} - \int_M u\,\text{div } X dV_g.$$

When M does not have a boundary, $\partial M = \emptyset$. So we have

$$\int_M \langle \text{grad } u, X \rangle_g dV_g = -\int_M u\,\text{div } X dV_g.$$

C.2 Normal coordinates

Riemannian normal coordinates. Let $x \in M$. There exist a neighborhood V of the origin in T_xM and a neighborhood U of x in M such that the exponential map $\exp_x : V \to U$ is a diffeomorphism. The set U is called a normal neighborhood of x. Given an orthonormal basis (z_i) of T_xM, there is a basis isomorphism from T_xM to $\mathbb{R}^d$. The exponential map can be combined with the basis isomorphism to get a smooth coordinate map $\varphi : U \to \mathbb{R}^d$. Such coordinates are called normal coordinates at x. Under normal coordinates, the coordinates of x is $0 \in \mathbb{R}^d$. For more details see for example Lee [2018, Chapter 5]

Cut locus and injectivity radius. Consider $v \in T_xM$ and let γ_v be the maximal geodesic starting at x with initial velocity v. Denote $t_{cut}(x,v) = \sup\{t > 0 : \text{ the restriction of } \gamma_v \text{ to } [0,t] \text{ is minimizing}\}$ The cut point of x along γ_v is $\gamma_v(t_{cut}(x,v))$ provided $t_{cut}(x,v) < \infty$. The cut locus of x is denoted as $\operatorname{Cut}(x) = \{q \in M : q \text{ is the cut point of } x \text{ along some geodesic.}\}$. The injectivity radius at x is the distance from x to its cut locus if the cut locus is nonempty, and infinite otherwise [Lee, 2018, Proposition 10.36]. When M is compact, the injectivity radius is positive [Lee, 2018, Lemma 6.16].

Theorem 13. *[Lee, 2018, Theorem 10.34] Let M be a complete, connected Riemannian manifold and $x \in M$. Then*

1. *The cut locus of x is a closed subset of M of measure zero.*
2. *The restriction of $\exp_x$ to $\overline{\operatorname{ID}(x)}$ is surjective.*
3. *The restriction of $\exp_x$ to $\operatorname{ID}(x)$ is a diffeomorphism onto $M \setminus \operatorname{Cut}(x)$.*

Here $\operatorname{ID}(x) = \{v \in T_xM : |v| < t_{cut}(x, \frac{v}{|v|})\}$ is the injectivity domain of x.

Then for any $p \in M$, under normal coordinates, for all well-behaved f, we have

$$\int_M f dV_g = \int_{M \setminus \operatorname{Cut}(p)} f dV_g = \int_{\varphi(M \setminus \operatorname{Cut}(p)) \subseteq \mathbb{R}^d} f(\varphi^{-1}(x)) \sqrt{\det(g)} dx.$$

C.3 Additional details for manifold of positive definite matrices

We briefly mention some properties of P_m. The inverse of the exponential map is globally defined and the cut locus of every point is empty. For symmetric matrix $S \in \mathbb{R}^{m \times m}$,

$$\begin{aligned}
\exp_X(tS) &= X^{1/2} \operatorname{Exp}(tX^{-1/2} S X^{-1/2}) X^{1/2}, \\
\gamma(t) &= X_1^{1/2} \operatorname{Exp}(t \operatorname{Log}(X_1^{-1/2} X_2 X_1^{-1/2})) X_1^{1/2} \text{ is a geodesic that connect } X_1, X_2, \\
d(X_1, X_2) &= \sqrt{\sum_{i=1}^m (\log(r_i))^2} \text{ with } r_i \text{ being eigenvalues of } X_1^{-1} X_2, \\
\exp_{X_1}^{-1}(X_2) &= \gamma'(0) = X_1^{1/2} \operatorname{Log}(X_1^{-1/2} X_2 X_1^{-1/2}) X_1^{1/2}.
\end{aligned}$$

We have the following fact.

Lemma 14. *Let $\phi(x) = d(x,y)^2$ with $y \in M$ being fixed. We have* $\operatorname{grad} \phi(x) = -2 \exp_x^{-1}(y)$.

D Proof of main Theorems

For a given ϕ, define the ϕ-divergence to be $\Phi_\pi(\rho) = \mathbb{E}_\pi[\phi(\frac{\rho}{\pi})]$. Define the following dissipation functional

$$D_\pi(\rho) := \mathbb{E}_\rho \left[\langle \operatorname{grad}(\phi' \circ \frac{\rho}{\pi}), \operatorname{grad} \log \frac{\rho}{\pi} \rangle \right].$$

We can now compute the time derivative of the ϕ-divergence along certain flow.

Let μ_t^X be the law of the continuous-time Langevin diffusion with target distribution $\pi^X \propto e^{-f(x)}$. That is, we have the following SDE, $dX_t = -\operatorname{grad} f(X_t) dt + \sqrt{2} dB_t$. Then, μ_t^X satisfies the following Fokker-Planck equation (see Lemma 25 for a proof).

$$\begin{aligned}
\frac{\partial}{\partial t} \mu_t^X &= \operatorname{div}(\mu_t^X \operatorname{grad} f(X_t) + \operatorname{grad} \mu_t^X) = \operatorname{div}(\operatorname{grad} \mu_t^X - \mu_t^X \frac{\operatorname{grad} \pi^X}{\pi^X}) \\
&= \operatorname{div}(\mu_t^X \operatorname{grad} \log \frac{\mu_t^X}{\pi^X}).
\end{aligned}$$

We now show that $D_{\pi^X}(\mu_t^X) = -\partial_t \Phi_{\pi^X}(\mu_t^X)$.

Lemma 15. *We have that*

$$D_{\pi^X}(\mu_t^X) := \mathbb{E}_{\mu_t^X}[\langle \operatorname{grad}(\phi' \circ \frac{\mu_t^X}{\pi^X}), \operatorname{grad} \log \frac{\mu_t^X}{\pi^X} \rangle] = -\partial_t \Phi_{\pi^X}(\mu_t^X).$$

Proof. [Proof of Lemma 15] By using the fact that $\frac{\partial}{\partial t}\mu_t^X = \operatorname{div}(\mu_t^X \operatorname{grad} \log \frac{\mu_t^X}{\pi^X})$, we have

$$\begin{aligned}\frac{\partial}{\partial t}\Phi_{\pi^X}(\mu_t^X) =& \frac{\partial}{\partial t}\int_M \pi^X \phi(\frac{\mu_t^X}{\pi^X}) dV_g(x) \\ =& \int_M \phi'(\frac{\mu_t^X}{\pi^X})\frac{\partial}{\partial t}\mu_t^X dV_g(x) = \int_M \Big(\phi'(\frac{\mu_t^X}{\pi^X})\Big)\Big(\operatorname{div}(\mu_t^X \operatorname{grad} \log \frac{\mu_t^X}{\pi^X})\Big) dV_g(x) \\ =& -\int_M \mu_t^X \langle \operatorname{grad} \phi' \circ \frac{\mu_t^X}{\pi^X}, \operatorname{grad} \log \frac{\mu_t^X}{\pi^X} \rangle dV_g(x),\end{aligned}$$

where in the last equality we used integration by parts. □

To get more intuition on the notion of ϕ-divergence and dissipation functional, consider $\phi(x) = x\log(x)$. We get KL divergence and fisher information:

$$\begin{aligned}&\Phi_\pi(\rho) = \mathbb{E}_\pi[\frac{\rho}{\pi}\log(\frac{\rho}{\pi})] = \mathbb{E}_\rho[\log(\frac{\rho}{\pi})] = H_\pi(\rho), \\ &D_\pi(\rho) = \mathbb{E}_\rho[\langle \operatorname{grad}(\phi' \circ \frac{\rho}{\pi}), \operatorname{grad} \log \frac{\rho}{\pi} \rangle] = \mathbb{E}_\rho[\| \operatorname{grad} \log \frac{\rho}{\pi}\|^2] = \mathbb{E}_\pi[\frac{\pi}{\rho}\| \operatorname{grad} \frac{\rho}{\pi}\|^2] = J_\pi(\rho).\end{aligned}$$

Our proof is now based on generalizing the proof in Chen et al. [2022] to the Riemannian setting. In Section D.1 we analyze the first step of proximal sampler by viewing it as simultaneous (forward) heat flow. In Section D.2 we analyze the second step of proximal sampler by viewing it as simultaneous backward flow. Combining the two steps together, we prove convergence of proximal sampler under LSI in Section D.3.

D.1 Forward step: simultaneous heat flow

We can first compute the time derivative of the ϕ-divergence along simultaneous heat flow.

Lemma 16. *Define Q_t to describe the forward heat flow. Let $\rho^X Q_t$ and $\pi^X Q_t$ evolve according to the simultaneous heat flow, satisfying*

$$\partial_t \rho^X Q_t = \frac{1}{2}\Delta(\rho^X Q_t) \quad \textit{and} \quad \partial_t \pi^X Q_t = \frac{1}{2}\Delta(\pi^X Q_t).$$

Then $\partial_t \Phi_{\pi^X Q_t}(\rho^X Q_t) = -\frac{1}{2} D_{\pi^X Q_t}(\rho^X Q_t)$.

Proof. [Proof of Lemma 16] Denote $\rho_t^X := \rho^X Q_t$ and $\pi_t^X := \pi^X Q_t$. Then, we have

$$\begin{aligned}2\frac{\partial}{\partial t}\Phi_{\pi_t^X}(\rho_t^X) =& 2\frac{\partial}{\partial t}\int_M \pi_t^X \phi(\frac{\rho_t^X}{\pi_t^X}) dV_g(x) \\ =& 2\int_M \phi(\frac{\rho_t^X}{\pi_t^X})\frac{\partial}{\partial t}\pi_t^X + \phi'(\frac{\rho_t^X}{\pi_t^X})\Big(\frac{\partial}{\partial t}\rho_t^X - (\frac{\partial}{\partial t}\pi_t^X)\frac{\rho_t^X}{\pi_t^X}\Big) dV_g(x).\end{aligned}$$

Recall that by construction,

$$\partial_t \rho_t^X = \frac{1}{2}\Delta\rho_t^X = \frac{1}{2}\operatorname{div}(\operatorname{grad} \rho_t^X) = \frac{1}{2}\operatorname{div}(\rho_t^X \operatorname{grad} \log \rho_t^X)$$

and $\partial_t \pi_t^X = \frac{1}{2}\Delta\pi_t^X = \frac{1}{2}\operatorname{div}(\pi_t^X \operatorname{grad} \log \pi_t^X)$. Hence, we get

$$\begin{aligned}2\frac{\partial}{\partial t}\Phi_{\pi_t^X}(\rho_t^X) =& 2\int_M \phi(\frac{\rho_t^X}{\pi_t^X})\frac{\partial}{\partial t}\pi_t^X + \phi'(\frac{\rho_t^X}{\pi_t^X})\Big(\frac{\partial}{\partial t}\rho_t^X - (\frac{\partial}{\partial t}\pi_t^X)\frac{\rho_t^X}{\pi_t^X}\Big) dV_g(x) \\ =& \int_M \phi(\frac{\rho_t^X}{\pi_t^X})\operatorname{div}(\pi_t^X \operatorname{grad} \log \pi_t^X)\end{aligned}$$

$$
\begin{aligned}
&+\phi'(\frac{\rho_t^X}{\pi_t^X})\Big(\operatorname{div}(\rho_t^X \operatorname{grad}\log\rho_t^X)-(\operatorname{div}(\pi_t^X \operatorname{grad}\log\pi_t^X))\frac{\rho_t^X}{\pi_t^X}\Big)dV_g(x)\\
=&\int_M -\Big\langle \operatorname{grad}\,\phi(\frac{\rho_t^X}{\pi_t^X}),\pi_t^X \operatorname{grad}\log\pi_t^X\Big\rangle\\
&-\Big\langle \operatorname{grad}\,\phi'(\frac{\rho_t^X}{\pi_t^X}),\rho_t^X \operatorname{grad}\log\rho_t^X\Big\rangle+\Big\langle \operatorname{grad}(\frac{\rho_t^X}{\pi_t^X}\phi'(\frac{\rho_t^X}{\pi_t^X})),\pi_t^X \operatorname{grad}\log\pi_t^X\Big\rangle dV_g(x)\\
=&\int_M -\Big\langle \operatorname{grad}\,\phi(\frac{\rho_t^X}{\pi_t^X}),\pi_t^X \operatorname{grad}\log\pi_t^X\Big\rangle-\Big\langle \operatorname{grad}\,\phi'(\frac{\rho_t^X}{\pi_t^X}),\rho_t^X \operatorname{grad}\log\rho_t^X\Big\rangle\\
&+\Big\langle \phi'(\frac{\rho_t^X}{\pi_t^X})\operatorname{grad}\,\frac{\rho_t^X}{\pi_t^X},\pi_t^X \operatorname{grad}\log\pi_t^X\Big\rangle+\Big\langle \frac{\rho_t^X}{\pi_t^X}\operatorname{grad}\,\phi'(\frac{\rho_t^X}{\pi_t^X}),\pi_t^X \operatorname{grad}\log\pi_t^X\Big\rangle dV_g(x).
\end{aligned}
$$

Now, notice that

$$
\Big\langle \operatorname{grad}\,\phi(\frac{\rho_t^X}{\pi_t^X}),\pi_t^X \operatorname{grad}\log\pi_t^X\Big\rangle=\Big\langle \phi'(\frac{\rho_t^X}{\pi_t^X})\operatorname{grad}\,\frac{\rho_t^X}{\pi_t^X},\pi_t^X \operatorname{grad}\log\pi_t^X\Big\rangle.
$$

So we get

$$
\begin{aligned}
2\frac{\partial}{\partial t}\Phi_{\pi_t^X}(\rho_t^X)&=\int_M\Big\langle \frac{\rho_t^X}{\pi_t^X}\operatorname{grad}\,\phi'(\frac{\rho_t^X}{\pi_t^X}),\pi_t^X \operatorname{grad}\log\pi_t^X\Big\rangle-\Big\langle \operatorname{grad}\,\phi'(\frac{\rho_t^X}{\pi_t^X}),\rho_t^X \operatorname{grad}\log\rho_t^X\Big\rangle dV_g(x)\\
&=-\int_M\rho_t^X\Big\langle \operatorname{grad}\,\phi'(\frac{\rho_t^X}{\pi_t^X}),\operatorname{grad}\log\frac{\rho_t^X}{\pi_t^X}\Big\rangle dV_g(x)\\
&=-\mathbb{E}_{\rho_t^X}\Big\langle \operatorname{grad}(\phi'\circ\frac{\rho_t^X}{\pi_t^X}),\operatorname{grad}\log\frac{\rho_t^X}{\pi_t^X}\Big\rangle=-D_{\pi_t^X}(\rho_t^X).
\end{aligned}
$$

□

D.2 Backward step: simultaneous backward flow

We leverage the following result.

Theorem 17 (Theorem 3.1 in De Bortoli et al. [2022])**.** *For a SDE $dX_t=b(X_t)dt+dB_t$, let p_t denote the distribution of X_t. Denote $Y_t=X_{T-t}, t\in[0,T]$ to be the time-reversed diffusion. We have that $dY_t=(-b(Y_t)+\operatorname{grad}\log p_{T-t}(Y_t))dt+dB_t$.*

Note that the time reversal can be understood as (Y_T,Y_0) has the same distribution as (X_0,X_T).

Recall that $\nu(t,x,y)$ is the density of manifold Brownian motion starting from x with time t and evaluated at y, and that $\pi(x,y)=\pi^X(x)\nu(\eta,x,y)$. We denote $\pi^Y=\pi^XQ_\eta$ to be the Y-marginal of $\pi(x,y)$. Let $\pi_t:=\pi^XQ_t$. Consider the forward process $dX_t=dB_t$ with $X_0\sim\pi^X$. We know that the time-reversed process satisfies $dY_t=\operatorname{grad}\log\pi_{\eta-t}(Y_t)dt+dB_t$.

Define Q_t^- as follows. Given ρ^Y, set $\rho^YQ_t^-$ to be the law at time t, of the solution of the time-reversed SDE (with $T=\eta$). Thus if $Y_0\sim\rho^Y$, we get $X_T\sim\rho^Y$. By Bayes theorem $X_0\sim\int_M\pi^{X|Y}(x|y)d\rho^Y(y)$, hence $Y_T\sim\int_M\pi^{X|Y}(x|y)d\rho^Y(y)$. For the channel Q_t^-, we have

1. Q_0^- is the identity channel.
2. Given input ρ^Y, the output at time η is $\rho^YQ_\eta^-(x)=\int_M\pi^{X|Y}(x|y)d\rho^Y(y)$.
3. $\pi^YQ_t^-=\pi^XQ_{\eta-t}$.

Thus we see that the RHK step of proximal sampler can be viewed as going along the time reversed process. We now have the following result.

Lemma 18. *For the time-reversed process, we have*

$$
\partial_t\Phi_{\pi^YQ_t^-}(\rho^YQ_t^-)=-\frac{1}{2}D_{\pi^YQ_t^-}(\rho^YQ_t^-).
$$

Proof. [Proof of Lemma 18] Denote $\pi_t^- = \pi^Y Q_t^-$ and $\rho_t^- = \rho^Y Q_t^-$. The Fokker-Planck equation is

$$\partial_t \pi_t^- = -\operatorname{div}(\pi_t^- \operatorname{grad} \log \pi_t^-) + \frac{1}{2}\Delta \pi_t^- = -\frac{1}{2}\Delta \pi_t^-,$$
$$\partial_t \rho_t^- = -\operatorname{div}(\rho_t^- \operatorname{grad} \log \pi_t^-) + \frac{1}{2}\Delta \rho_t^- = \operatorname{div}(\rho_t^- \operatorname{grad} \log \frac{\rho_t^-}{\pi_t^-}) - \frac{1}{2}\Delta \rho_t^-.$$

Hence

$$\begin{aligned}
2\partial_t \Phi_{\pi_t^-}(\rho_t^-) =& 2\int_M \phi(\frac{\rho_t^-}{\pi_t^-})\frac{\partial}{\partial t}\pi_t^- + \phi'(\frac{\rho_t^-}{\pi_t^-})\Big(\frac{\partial}{\partial t}\rho_t^- - (\frac{\partial}{\partial t}\pi_t^-)\frac{\rho_t^-}{\pi_t^-}\Big) dV_g(x) \\
=& \int_M -\phi(\frac{\rho_t^-}{\pi_t^-})\Delta \pi_t^- \\
& + \phi'(\frac{\rho_t^-}{\pi_t^-})\Big(2\operatorname{div}(\rho_t^- \operatorname{grad} \log \frac{\rho_t^-}{\pi_t^-}) - \Delta \rho_t^- + \frac{\rho_t^-}{\pi_t^-}\Delta \pi_t^-\Big) dV_g(x) \\
=& 2\int_M \phi'(\frac{\rho_t^-}{\pi_t^-})\operatorname{div}(\rho_t^- \operatorname{grad} \log \frac{\rho_t^-}{\pi_t^-}) dV_g(x) \\
& + \int_M -\phi(\frac{\rho_t^-}{\pi_t^-})\Delta \pi_t^- - \phi'(\frac{\rho_t^-}{\pi_t^-})\Delta \rho_t^- + \phi'(\frac{\rho_t^-}{\pi_t^-})\frac{\rho_t^-}{\pi_t^-}\Delta \pi_t^- dV_g(x) \\
=& 2\int_M \phi'(\frac{\rho_t^-}{\pi_t^-})\operatorname{div}(\rho_t^- \operatorname{grad} \log \frac{\rho_t^-}{\pi_t^-}) dV_g(x) + D_{\pi_t^-}(\rho_t^-) \\
=& -2D_{\pi_t^-}(\rho_t^-) + D_{\pi_t^-}(\rho_t^-) = -D_{\pi_t^-}(\rho_t^-),
\end{aligned}$$

where we used the same steps as as in the proof Lemma 16 to obtain

$$\int_M -\phi(\frac{\rho_t^-}{\pi_t^-})\Delta \pi_t^- - \phi'(\frac{\rho_t^-}{\pi_t^-})\Delta \rho_t^- + \phi'(\frac{\rho_t^-}{\pi_t^-})\frac{\rho_t^-}{\pi_t^-}\Delta \pi_t^- dV_g(x) = D_{\pi_t^-}(\rho_t^-),$$

and used integration by parts, to obtain

$$\begin{aligned}
& 2\int_M \phi'(\frac{\rho_t^-}{\pi_t^-})\operatorname{div}(\rho_t^- \operatorname{grad} \log \frac{\rho_t^-}{\pi_t^-}) dV_g(x) = -2\int_M \langle \operatorname{grad} \phi'(\frac{\rho_t^-}{\pi_t^-}), \rho_t^- \operatorname{grad} \log \frac{\rho_t^-}{\pi_t^-}\rangle dV_g(x) \\
=& -2\mathbb{E}_{\rho_t^-}[\langle \operatorname{grad} \phi'(\frac{\rho_t^-}{\pi_t^-}), \operatorname{grad} \log \frac{\rho_t^-}{\pi_t^-}\rangle] = -2D_{\pi_t^-}(\rho_t^-).
\end{aligned}$$

□

D.3 Convergence under LSI

Now we prove the main theorem.

Proof. [Proof of Theorem 6 item 1] We first prove the theorem assuming curvature is non-negative. For the general case, we only need to replace the LSI constant α_t, α_t^-.

1. **The forward step.** We know $\pi^X Q_t$ satisfies LSI with $\alpha_t := \frac{1}{t + \frac{1}{\alpha}}$. Using Lemma 16, we have

$$\partial_t H_{\pi^X Q_t}(\rho_0^X Q_t) = -\frac{1}{2} J_{\pi^X Q_t}(\rho_0^X Q_t) \le -\alpha_t H_{\pi^X Q_t}(\rho_0^X Q_t).$$

   This implies $H_{\pi^X Q_t}(\rho_0^X Q_t) \le e^{-A_t} H_{\pi^X}(\rho_0^X)$ where, $A_t = \int_0^t \alpha_s ds = \log(1 + t\alpha)$. We also have $e^{-A_t} = (1 + t\alpha)^{-1}$. As a result,

$$H_{\pi^X Q_\eta}(\rho_0^X Q_\eta) \le \frac{H_{\pi^X}(\rho_0^X)}{1 + \eta\alpha}.$$

2. **The backward step.** Using Lemma 18, we have

$$\partial_t H_{\pi^Y Q_t^-}(\rho_0^Y Q_t^-) = -\frac{1}{2} J_{\pi^Y Q_t^-}(\rho_0^Y Q_t^-) \le -\alpha_t^- H_{\pi^Y Q_t^-}(\rho_0^Y Q_t^-).$$

Since $\pi^Y Q_t^- = \pi^X Q_{\eta-t}$, we know the LSI constant for $\pi^Y Q_t^-$ is $\alpha_t^- := \frac{1}{(\eta-t)+\frac{1}{\alpha}}$. Same as in step 1, we get $A_\eta^- = \int_0^\eta \alpha_t^- dt = \log(1+\alpha\eta)$. As a result,

$$H_{\pi^Y Q_\eta^-}(\rho_0^Y Q_\eta^-) \le \frac{H_{\pi^Y}(\rho_0^Y)}{1+t\alpha}.$$

3. **Putting together.** We have $\pi^Y = \pi^X Q_\eta$, $\rho_0^Y = \rho_0^X Q_\eta$. Denote $\rho_1^X = \rho_0^Y Q_\eta^-$, we get

$$H_{\pi^X}(\rho_1^X) = H_{\pi^Y Q_\eta^-}(\rho_0^Y Q_\eta^-) \le \frac{H_{\pi^Y}(\rho_0^Y)}{(1+t\alpha)} = \frac{H_{\pi^X Q_\eta}(\rho_0^X Q_\eta)}{(1+t\alpha)} \le \frac{H_{\pi^X}(\rho_0^X)}{(1+t\alpha)^2}.$$

4. **Negative curvature.** For negative curvature, we use α_t as in Proposition 31 (the value to be integrated is $\frac{\kappa}{1-e^{-\kappa t}+\kappa d_0 e^{-\kappa t}}$ where $\frac{1}{\alpha} := d_0$). We compute the integral

$$\int_0^t \frac{1}{((\frac{1}{\alpha}-\frac{1}{\kappa})e^{-x}+\frac{1}{\kappa})} dx = \log(\alpha(e^{\kappa t}-1)+\kappa) - \log(\kappa) = \log(\frac{\alpha(e^{\kappa t}-1)+\kappa}{\kappa}).$$

Hence we have $H_{\pi^X}(\rho_k^X) \le H_{\pi^X}(\rho_0^X)(\frac{\kappa}{\alpha(e^{\kappa\eta}-1)+\kappa})^{2k}$.

Observe that in general, for $x \in [0,1]$ we have that $1-\frac{x}{2} \ge e^{-x}$. Thus for $\eta < \frac{1}{|\kappa|}$, we have $|\kappa|\eta < 1$, hence $1-\frac{|\kappa|}{2}\eta \ge e^{-|\kappa|\eta}$. This implies $\frac{1-e^{-|\kappa|\eta}}{|\kappa|} \ge \frac{1}{2}\eta$. On the other hand, we have $1-x \le e^{-x}$, which implies $\frac{1-e^{-|\kappa|\eta}}{|\kappa|} \le \eta$. So we have $\frac{\alpha(e^{\kappa\eta}-1)+\kappa}{\kappa} = 1+\alpha\frac{1-e^{-|\kappa|\eta}}{|\kappa|} = \Theta(1+\alpha\eta)$.

To summarize, we have that when $\kappa < 0$,

$$\frac{\kappa}{\alpha(e^{\kappa\eta}-1)+\kappa} = \frac{1}{1+\alpha\frac{1-e^{-|\kappa|\eta}}{|\kappa|}} \le \frac{1}{1+\frac{\alpha}{2}\eta}$$

which implies $H_{\pi^X}(\rho_k^X) \le H_{\pi^X}(\rho_0^X)(\frac{1}{1+\frac{\alpha}{2}\eta})^{2k}$

□

D.4 Convergence under Poincaré inequality

In this subsection, we extend Theorem 6 item 1 from LSI to PI, under χ^2 divergence. We follow exactly the same strategy, but we first discuss how to modify the proof from the LSI setting to PI setting. Recall that if π satisfies Poincaré inequality with parameter α, we have

$$\mathrm{Var}_\pi(\frac{d\mu}{d\pi}) \le \frac{1}{\alpha}\mathbb{E}[\|\operatorname{grad}\frac{d\mu}{d\pi}\|^2]$$

Note that χ^2 divergence is a ϕ-divergence with $\phi(x) = (x-1)^2$, which allows us have $\mathrm{Var}_\pi(\frac{d\mu}{d\pi}) = \mathbb{E}_\pi[(\frac{d\mu}{d\pi})^2] - \mathbb{E}_\pi[\frac{d\mu}{d\pi}]^2 = \chi^2_\pi(\mu)$. Also, recall that we defined in Appendix D the dissipation of ϕ-divergence. By definition, the dissipation of χ^2 divergence is $D_\pi(\mu) = 2\mathbb{E}_\rho[\langle\operatorname{grad}(\frac{\mu}{\pi}-1), \operatorname{grad}\log\frac{\mu}{\pi}\rangle] = 2\mathbb{E}_\pi[\|\operatorname{grad}\frac{d\mu}{d\pi}\|^2]$. Hence we can interpret the Poincaré inequality as

$$\chi^2_\pi(\mu) \le \frac{1}{2\alpha} D_\pi(\mu)$$

From this perspective, to follow the proof for LSI setting, we need to know whether the Poincaré inequality constant is also preserved along heat propagation, in the same way as LSI constant. The answer is yes, see Appendix H.3 Proposition 32.

With this, we can prove Theorem 6 item 2 following exactly the same procedure as Theorem 6 item 1.

Proof. [Proof of Theorem 6 item 2] We first assume curvature is non-negative, then discuss the situation that curvature is negative.

1. **The forward step.** We know $\pi^X Q_t$ satisfies PI with $\alpha_t := \frac{1}{t+\frac{1}{\alpha}}$. Using Lemma 16 and Proposition 32, we have

$$\partial_t \chi^2_{\pi^X Q_t}(\rho_0^X Q_t) = -\frac{1}{2} D_{\pi^X Q_t}(\rho_0^X Q_t) \leq -\alpha_t \chi^2_{\pi^X Q_t}(\rho_0^X Q_t).$$

This implies $\chi^2_{\pi^X Q_t}(\rho_0^X Q_t) \leq e^{-A_t} H_{\pi^X}(\rho_0^X)$ where, $A_t = \int_0^t \alpha_s ds = \log(1+t\alpha)$. We also have $e^{-A_t} = (1+t\alpha)^{-1}$. As a result,

$$\chi^2_{\pi^X Q_\eta}(\rho_0^X Q_\eta) \leq \frac{\chi^2_{\pi^X}(\rho_0^X)}{1+\eta\alpha}.$$

2. **The backward step.** Using Lemma 18 and Proposition 32, we have

$$\partial_t \chi^2_{\pi^Y Q_t^-}(\rho_0^Y Q_t^-) = -\frac{1}{2} D_{\pi^Y Q_t^-}(\rho_0^Y Q_t^-) \leq -\alpha_t^- \chi^2_{\pi^Y Q_t^-}(\rho_0^Y Q_t^-).$$

Since $\pi^Y Q_t^- = \pi^X Q_{\eta - t}$, we know the LSI constant for $\pi^Y Q_t^-$ is $\alpha_t^- := \frac{1}{(\eta - t) + \frac{1}{\alpha}}$. Same as in step 1, we get $A_\eta^- = \int_0^\eta \alpha_t^- dt = \log(1+\alpha\eta)$. As a result,

$$\chi^2_{\pi^Y Q_\eta^-}(\rho_0^Y Q_\eta^-) \leq \frac{\chi^2_{\pi^Y}(\rho_0^Y)}{1+t\alpha}.$$

3. **Putting together.** We have $\pi^Y = \pi^X Q_\eta$, $\rho_0^Y = \rho_0^X Q_\eta$. Denote $\rho_1^X = \rho_0^Y Q_\eta^-$, we get

$$\chi^2_{\pi^X}(\rho_1^X) = \chi^2_{\pi^Y Q_\eta^-}(\rho_0^Y Q_\eta^-) \leq \frac{\chi^2_{\pi^Y}(\rho_0^Y)}{(1+t\alpha)} = \frac{\chi^2_{\pi^X Q_\eta}(\rho_0^X Q_\eta)}{(1+t\alpha)} \leq \frac{\chi^2_{\pi^X}(\rho_0^X)}{(1+t\alpha)^2}.$$

4. **Negative curvature.** For negative curvature, we use α_t as in Proposition 31 (the value to be integrated is $\frac{\kappa}{1-e^{-\kappa t}+\kappa d_0 e^{-\kappa t}}$ where $\frac{1}{\alpha} := d_0$). We compute the integral

$$\int_0^t \frac{1}{((\frac{1}{\alpha} - \frac{1}{\kappa})e^{-x} + \frac{1}{\kappa})} dx = \log(\alpha(e^{\kappa t} - 1) + \kappa) - \log(\kappa) = \log(\frac{\alpha(e^{\kappa t} - 1) + \kappa}{\kappa}).$$

Hence we have $H_{\pi^X}(\rho_k^X) \leq H_{\pi^X}(\rho_0^X)(\frac{\kappa}{\alpha(e^{\kappa\eta}-1)+\kappa})^{2k}$.

Observe that in general, for $x \in [0,1]$ we have that $1 - \frac{x}{2} \geq e^{-x}$. Thus for $\eta < \frac{1}{|\kappa|}$, we have $|\kappa|\eta < 1$, hence $1 - \frac{|\kappa|}{2}\eta \geq e^{-|\kappa|\eta}$. This implies $\frac{1-e^{-|\kappa|\eta}}{|\kappa|} \geq \frac{1}{2}\eta$. On the other hand, we have $1 - x \leq e^{-x}$, which implies $\frac{1-e^{-|\kappa|\eta}}{|\kappa|} \leq \eta$. So we have $\frac{\alpha(e^{\kappa\eta}-1)+\kappa}{\kappa} = 1 + \alpha\frac{1-e^{-|\kappa|\eta}}{|\kappa|} = \Theta(1+\alpha\eta)$.

To summarize, we have that when $\kappa < 0$,

$$\frac{\kappa}{\alpha(e^{\kappa\eta}-1)+\kappa} = \frac{1}{1+\alpha\frac{1-e^{-|\kappa|\eta}}{|\kappa|}} \leq \frac{1}{1+\frac{\alpha}{2}\eta}$$

which implies $\chi^2_{\pi^X}(\rho_k^X) \leq \chi^2_{\pi^X}(\rho_0^X)(\frac{1}{1+\frac{\alpha}{2}\eta})^{2k}$

□

E Proof of Theorem 8

Recall that $\rho_k^X(x)$, $\rho_k^Y(y)$ denote the distribution generated by Algorithm 1, assuming exact Brownian motion and exact RHK. This notation is applied for all k. For practical implementation, using inexact RHK and inexact Brownian motion through all the iterations, we denote the corresponding distribution by $\tilde{\rho}_k^X(x)$, $\tilde{\rho}_k^Y(y)$ respectively.

Note that at iteration $k-1$, we are at distribution $\tilde{\rho}_{k-1}^X(x)$. Denote $\hat{\rho}_{k-1}^Y(y)$ to be the distribution obtained from $\tilde{\rho}_{k-1}^X(x)$ using exact Brownian motion. (Note that $\tilde{\rho}_{k-1}^Y(y)$ denote the distribution obtained from $\tilde{\rho}_{k-1}^X(x)$ using inexact Brownian motion).

We now prove Lemma 7.

Proof. [Proof of Lemma 7] Using triangle inequality, we have

$$\begin{aligned}&\|\rho_k^X(x)-\tilde{\rho}_k^X(x)\|_{TV}=\Big\|\int\rho_{k-1}^Y(y)\pi^{X|Y}(x|y)dy-\int\tilde{\rho}_{k-1}^Y(y)\hat{\pi}^{X|Y}(x|y)dy\Big\|_{TV}\\ \leq&\Big\|\int\rho_{k-1}^Y(y)(\pi^{X|Y}(x|y)-\hat{\pi}^{X|Y}(x|y))dy\Big\|_{TV}+\Big\|\int(\tilde{\rho}_{k-1}^Y(y)-\rho_{k-1}^Y(y))\hat{\pi}^{X|Y}(x|y)dy\Big\|_{TV}.\end{aligned}$$

The first part can be bounded by ζ_{RHK}:

$$\begin{aligned}&\Big\|\int\rho_{k-1}^Y(y)(\pi^{X|Y}(x|y)-\hat{\pi}^{X|Y}(x|y))dy\Big\|_{TV}\leq\int\rho_{k-1}^Y(y)\Big\|\pi^{X|Y}(x|y)-\hat{\pi}^{X|Y}(x|y)\Big\|_{TV}dy\\ \leq&\zeta_{\mathsf{RHK}}.\end{aligned}$$

For the second part, we have

$$\begin{aligned}&\Big\|\int(\tilde{\rho}_{k-1}^Y(y)-\rho_{k-1}^Y(y))\hat{\pi}^{X|Y}(x|y)dy\Big\|_{TV}\\ =&\frac{1}{2}\int\Big|\int(\tilde{\rho}_{k-1}^Y(y)-\rho_{k-1}^Y(y))\hat{\pi}^{X|Y}(x|y)dy\Big|dx\leq\frac{1}{2}\int\Big|\tilde{\rho}_{k-1}^Y(y)-\rho_{k-1}^Y(y)\Big|\int\hat{\pi}^{X|Y}(x|y)dxdy\\ =&\|\tilde{\rho}_{k-1}^Y(y)-\rho_{k-1}^Y(y)\|_{TV}\leq\|\tilde{\rho}_{k-1}^Y(y)-\hat{\rho}_{k-1}^Y(y)\|_{TV}+\|\hat{\rho}_{k-1}^Y(y)-\rho_{k-1}^Y(y)\|_{TV}\\ \leq&\zeta_{\mathsf{MBI}}+\|\tilde{\rho}_{k-1}^X(x)-\rho_{k-1}^X(x)\|_{TV}.\end{aligned}$$

Here, the last inequality follows from Lemma 37. Together, we have

$$\|\rho_k^X(x)-\tilde{\rho}_k^X(x)\|_{TV}\leq\zeta_{\mathsf{RHK}}+\zeta_{\mathsf{MBI}}+\|\tilde{\rho}_{k-1}^X(x)-\rho_{k-1}^X(x)\|_{TV}.$$

Iteratively applying this inequality and noting that $\|\tilde{\rho}_0^X(x)-\rho_0^X(x)\|_{TV}=0$, we obtain $\|\rho_k^X(x)-\tilde{\rho}_k^X(x)\|_{TV}\leq k(\zeta_{\mathsf{RHK}}+\zeta_{\mathsf{MBI}})$. □

Recall that Pinsker's inequality states $\|\mu-\nu\|_{TV}\leq\sqrt{\frac{1}{2}H_\nu(\mu)}$.

Proof. [Proof of Theorem 8 item 1]

For simplicity, we assume non-negative curvature. The negative curvature case follows from the same proof strategy. Using Pinsker's inequality, we have

$$\|\rho_k^X-\pi^X\|_{TV}\leq\sqrt{\frac{1}{2}H_{\pi^X}(\rho_k^X)}\leq\sqrt{\frac{1}{2}\frac{H_{\pi^X}(\rho_0^X)}{(1+\eta\alpha)^{2k}}}\leq\frac{1}{2}\varepsilon.$$

We want to bound $\|\rho_k^X-\pi^X\|_{TV}\leq\frac{1}{2}\varepsilon$. It suffices to have $\frac{H_{\pi^X}(\rho_0^X)}{(1+\eta\alpha)^{2k}}\leq\frac{1}{2}\varepsilon^2$. Hence we need $\log(\frac{2H_{\pi^X}(\rho_0^X)}{\varepsilon^2})\leq 2k\log(1+\eta\alpha)$, i.e., $k=\mathcal{O}\left(\frac{\log\frac{H_{\pi^X}(\rho_0^X)}{\varepsilon^2}}{\log(1+\eta\alpha)}\right)$.

For small step size η, we have $\frac{1}{\log(1+\eta\alpha)}=\mathcal{O}(\frac{1}{\eta\alpha})$. Hence $k=\mathcal{O}\left(\frac{1}{\eta\alpha}\log\frac{H_{\pi^X}(\rho_0^X)}{\varepsilon^2}\right)=\tilde{\mathcal{O}}(\frac{1}{\eta}\log\frac{1}{\varepsilon})$. Recall that by assumption, $\frac{1}{\eta}=\tilde{\mathcal{O}}(\log\frac{1}{\zeta})$. We pick $\zeta=\frac{\varepsilon}{\log^2\frac{1}{\varepsilon}}$ and consequently $\frac{1}{\eta}=\tilde{\mathcal{O}}(\log\frac{\log^2\frac{1}{\varepsilon}}{\varepsilon})=\tilde{\mathcal{O}}(\log\frac{1}{\varepsilon}+2\log\log\frac{1}{\varepsilon})=\tilde{\mathcal{O}}(\log\frac{1}{\varepsilon})$. It follows that

$$k=\tilde{\mathcal{O}}(\frac{1}{\eta}\log\frac{1}{\varepsilon})=\tilde{\mathcal{O}}(\log^2\frac{1}{\varepsilon}).$$

The result then follows from triangle inequality:

$$\|\tilde{\rho}_k^X-\pi^X\|_{TV}\leq\|\tilde{\rho}_k^X-\rho_k^X\|_{TV}+\|\rho_k^X-\pi^X\|_{TV}\leq k(\zeta_{\mathsf{RHK}}+\zeta_{\mathsf{MBI}})+\frac{1}{2}\varepsilon=\tilde{\mathcal{O}}(\varepsilon)$$

where $k\zeta=\tilde{\mathcal{O}}(\frac{\varepsilon}{\log^2\frac{1}{\varepsilon}}\log^2\frac{1}{\varepsilon})=\tilde{\mathcal{O}}(\varepsilon)$.

□

Proof. [Proof of Theorem 8 item 2] For simplicity, we assume non-negative curvature. The negative curvature case follows from the same proof strategy. We know that for two probability measures π, ρ we have $\log(1+\chi^2_\pi(\rho)) \geq H_\pi(\rho) \geq 2\|\pi-\rho\|^2_{TV}$ where the first inequality is [Nishiyama and Sason, 2020, equation 16] and the second inequality is Pinsker's inequality.

Hence, we have

$$\|\rho_k^X - \pi^X\|_{TV} \leq \sqrt{\frac{1}{2}H_{\pi^X}(\rho_k^X)} \leq \sqrt{\frac{1}{2}\log(1+\chi^2_{\pi^X}(\rho_k^X))} \leq \sqrt{\frac{1}{2}\log\left(1+\frac{\chi^2_{\pi^X}(\rho_0^X)}{(1+\eta\frac{\alpha}{2})^{2k}}\right)}.$$

We want to bound $\|\rho_k^X - \pi^X\|_{TV} \leq \frac{1}{2}\varepsilon$. It suffices to have $\log\left(1+\frac{\chi^2_{\pi^X}(\rho_0^X)}{(1+\eta\frac{\alpha}{2})^{2k}}\right) \leq \frac{1}{2}\varepsilon^2$. We need $\chi^2_{\pi^X}(\rho_0^X) \leq (\exp(\frac{1}{2}\varepsilon^2)-1)(1+\eta\frac{\alpha}{2})^{2k}$, that is,

$$k \geq \frac{1}{2\log(1+\eta\frac{\alpha}{2})}\log\frac{\chi^2_{\pi^X}(\rho_0^X)}{(\exp(\frac{1}{2}\varepsilon^2)-1)} = \mathcal{O}(\frac{1}{\log(1+\eta\frac{\alpha}{2})}\log\frac{\chi^2_{\pi^X}(\rho_0^X)}{\varepsilon^2})$$

For small step size η, we have $\frac{1}{\log(1+\eta\alpha/2)} = \mathcal{O}(\frac{1}{\eta\alpha})$. Hence $k = \mathcal{O}\left(\frac{1}{\eta\alpha}\log\frac{\chi^2_{\pi^X}(\rho_0^X)}{\varepsilon^2}\right)$. Recall that by assumption, $\frac{1}{\eta} = \tilde{\mathcal{O}}(\log\frac{1}{\zeta})$. We pick $\zeta = \frac{\varepsilon}{\log^2\frac{1}{\varepsilon}}$ and consequently $\frac{1}{\eta} = \tilde{\mathcal{O}}(\log\frac{\log^2\frac{1}{\varepsilon}}{\varepsilon}) = \tilde{\mathcal{O}}(\log\frac{1}{\varepsilon}+2\log\log\frac{1}{\varepsilon}) = \tilde{\mathcal{O}}(\log\frac{1}{\varepsilon})$. It follows that

$$k = \tilde{\mathcal{O}}\left(\frac{1}{\eta}\log\frac{\chi^2_{\pi^X}(\rho_0^X)}{\varepsilon^2}\right) = \tilde{\mathcal{O}}\left(\log\frac{1}{\varepsilon}\log\frac{\chi^2_{\pi^X}(\rho_0^X)}{\varepsilon^2}\right).$$

The result then follows from triangle inequality:

$$\|\tilde{\rho}_k^X - \pi^X\|_{TV} \leq \|\tilde{\rho}_k^X - \rho_k^X\|_{TV} + \|\rho_k^X - \pi^X\|_{TV} \leq k(\zeta_{\mathsf{RHK}} + \zeta_{\mathsf{MBI}}) + \frac{1}{2}\varepsilon = \tilde{\mathcal{O}}(\varepsilon)$$

where $k\zeta = \tilde{\mathcal{O}}(\varepsilon)$.

□

F Verification of Assumption 1

In this section, we consider implementing inexact oracles through the truncation method. Recall that we assume M is a compact manifold, which is a homogeneous space.

We use $\hat{\pi}^{Y|X}, \hat{\pi}^{X|Y}$ to denote the output of MBI oracle and RHK when rejection sampling is exact. More precisely, since we use the truncated series to approximate heat kernel, we have $\hat{\pi}^{Y|X} \propto \nu_l(\eta, x, y)$ and $\hat{\pi}^{X|Y} \propto e^{-f(x)}\nu_l(\eta, x, y)$. When rejection sampling is not exact, i.e., there exists $z \in M$ s.t. $V(z) > 1$, we denote the output to be $\overline{\pi}^{Y|X}, \overline{\pi}^{X|Y}$ for inexact Brownian motion and inexact RHK, respectively.

In subsection F.1, we prove Proposition 9, i.e., $\|\hat{\pi}^{X|Y} - \pi^{X|Y}\|_{TV} = \tilde{\mathcal{O}}(\zeta)$ and $\|\hat{\pi}^{Y|X} - \pi^{Y|X}\|_{TV} = \tilde{\mathcal{O}}(\zeta)$ with $\zeta = \frac{\varepsilon}{\log^2\frac{1}{\varepsilon}}$. Then we know $\hat{\pi}^{Y|X}$ and $\hat{\pi}^{X|Y}$ satisfy Assumption 1.

In subsection I.1, we consider a more general setting, where the acceptance rate is allowed to exceed 1 at some unimportant region. We show that on $\mathcal{S}^d$, for certain choices of parameters, $\|\hat{\pi}^{X|Y} - \overline{\pi}^{X|Y}\|_{TV} = \tilde{\mathcal{O}}(\zeta)$ and $\|\hat{\pi}^{Y|X} - \overline{\pi}^{Y|X}\|_{TV} = \tilde{\mathcal{O}}(\zeta)$. This means that allowing the acceptance rate to exceed 1 in unimportant regions would not cause a significant bias for rejection sampling. It then follows from triangle inequality that $\overline{\pi}^{Y|X}$ and $\overline{\pi}^{X|Y}$ satisfy Assumption 1.

F.1 Exact rejection sampling

We prove Proposition 9, i.e., verify that Assumption 1 is satisfied with $\zeta = \frac{\varepsilon}{\log^2\frac{1}{\varepsilon}}$ as required in Theorem 8.

F.1.1 Analysis in total variation distance

The first step is to bound the total variation distance, under the assumption that heat kernel evaluation is of high accuracy. We consider the following characterization of total variation distance (see Lemma 36):

$$\|\rho_1 - \rho_2\|_{TV} = \frac{1}{2}\int_M |\rho_1(x) - \rho_2(x)| dV_g(x).$$

Proposition 19. *Let M be a compact manifold. Let ζ be the desired accuracy. Assume for all $y \in M$ we have $\int_M |\nu(\eta, x, y) - \nu_l(\eta, x, y)| dV_g(x) = \tilde{\mathcal{O}}(\zeta)$ and for all $x \in M$ we have $\int_M |\nu(\eta, x, y) - \nu_l(\eta, x, y)| dV_g(y) = \tilde{\mathcal{O}}(\zeta)$. Then $\|\hat{\pi}^{X|Y} - \pi^{X|Y}\|_{TV} = \tilde{\mathcal{O}}(\zeta)$ and $\|\hat{\pi}^{Y|X} - \pi^{Y|X}\| = \tilde{\mathcal{O}}(\zeta)$.*

Proof. [Proof of Proposition 19]

Step 1. Note that $A_1 := \sup_{x\in M} e^{-f(x)}$, $A_2 := \inf_{x\in M} e^{-f(x)}$ are positive constants independent of t. Denote $Z_1 = \int_M e^{-f(x)}\nu_l(\eta, x, y)dV_g(x)$ and $Z_2 = \int_M e^{-f(x)}\nu(\eta, x, y)dV_g(x)$. We know

$$\begin{aligned}
|Z_2 - Z_1| &= |\int_M e^{-f(x)}\nu(\eta, x, y) - e^{-f(x)}\nu_l(\eta, x, y)dV_g(x)| \\
&\leq A_1 \int_M |\nu(\eta, x, y) - \nu_l(\eta, x, y)| dV_g(x) = \tilde{\mathcal{O}}(\zeta).
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\|\hat{\pi}^{X|Y} - \pi^{X|Y}\|_{TV} &\leq \frac{1}{2}\int_M |\frac{e^{-f(x)}\nu_l(\eta, x, y)}{\int_M e^{-f(x)}\nu_l(\eta, x, y)dV_g(x)} - \frac{e^{-f(x)}\nu(\eta, x, y)}{\int_M e^{-f(x)}\nu(\eta, x, y)dV_g(x)}| dV_g(x) \\
&\leq \frac{1}{2}\int_M \frac{|Z_2 e^{-f(x)}\nu_l(\eta, x, y) - Z_1 e^{-f(x)}\nu(\eta, x, y)|}{Z_1 Z_2} dV_g(x) \\
&\leq \int_M \frac{\min\{Z_1, Z_2\} \cdot |e^{-f(x)}\nu_l(\eta, x, y) - e^{-f(x)}\nu(\eta, x, y)|}{2Z_1 Z_2} dV_g(x) \\
&\quad + \int_M \frac{|Z_2 - Z_1| \cdot \max\{e^{-f(x)}\nu(\eta, x, y), e^{-f(x)}\nu_l(\eta, x, y)\}}{2Z_1 Z_2} dV_g(x) \\
&\leq \tilde{\mathcal{O}}(\zeta) + \tilde{\mathcal{O}}(\int_M |Z_2 - Z_1| \cdot \max\{\nu(\eta, x, y), \nu_l(\eta, x, y)\} dV_g(x)) = \tilde{\mathcal{O}}(\zeta),
\end{aligned}$$

where by Lemma 21, we obtain $\frac{\min\{Z_1, Z_2\}}{2Z_1 Z_2} = \tilde{\mathcal{O}}(1)$ and

$$\int_M \frac{\max\{e^{-f(x)}\nu(\eta, x, y), e^{-f(x)}\nu_k(\eta, x, y)\}}{2Z_1 Z_2} dV_g(x) = \tilde{\mathcal{O}}(1).$$

Step 2. Denote $Z_l = \int_M \nu_l(\eta, x, y)dV_g(y)$ to be the normalizaing constant for ν_l. Since ν is the heat kernel, we simply have $\int_M \nu(\eta, x, y)dV_g(y) = 1$. It holds that

$$\hat{\pi}^{Y|X} = \frac{\nu_l(\eta, x, y)}{\int_M \nu_l(\eta, x, y)dV_g(y)} \quad \text{and} \quad \pi^{Y|X} = \nu(\eta, x, y).$$

Then,

$$\begin{aligned}
\|\hat{\pi}^{Y|X} - \pi^{Y|X}\|_{TV} &\leq \frac{1}{2}\int_M |\frac{\nu_l(\eta, x, y)}{\int_M \nu_l(\eta, x, y)dV_g(y)} - \nu(\eta, x, y)| dV_g(y) \\
&\leq \frac{1}{2}\int_M \frac{|\nu_l(\eta, x, y) - Z_l\nu(\eta, x, y)|}{Z_l} dV_g(y) \\
&\leq \int_M \frac{\min\{Z_l, 1\} \cdot |\nu_l(\eta, x, y) - \nu(\eta, x, y)|}{2Z_l} dV_g(y) \\
&\quad + \int_M \frac{|1 - Z_l| \cdot \max\{\nu(\eta, x, y), \nu_l(\eta, x, y)\}}{2Z_l} dV_g(y)
\end{aligned}$$

$$= \tilde{\mathcal{O}}(\zeta).$$

□

Theorem 20 (Theorem 5.3.4 in Hsu [2002])**.** *Let M be a compact Riemannian manifold. There exist positive constants C_1, C_2 such that for all $(t, x, y) \in (0, 1) \times M \times M$,*

$$\frac{C_1}{t^{d/2}} e^{-\frac{d(x,y)^2}{2t}} \leq \nu(t, x, y) \leq \frac{C_2}{t^{(2d-1)/2}} e^{-\frac{d(x,y)^2}{2t}}.$$

Lemma 21. *We have $1/\int_M e^{-f(x)}\nu(\eta, x, y)dV_g(x) = \tilde{\mathcal{O}}(1)$.*

Proof. [Proof of Lemma 21] Using lower bound of heat kernel from Theorem 20, we have

$$\begin{aligned}
&\int_M e^{-f(x)}\nu(\eta, x, y)dV_g(x) \\
\geq& A_2 \int_M \frac{C_1}{\eta^{d/2}} \exp(-\frac{d(x,y)^2}{2\eta})dV_g(x) = \frac{A_2 C_1}{\eta^{d/2}} \int_M \exp(-\frac{d(x,y)^2}{2\eta})dV_g(x) \\
\geq& \frac{A_2 C_1}{\eta^{d/2}} \frac{\eta^{d/2}}{C_4} = \frac{A_2 C_1}{C_4}.
\end{aligned}$$

Hence

$$1/\int_M e^{-f(x)}\nu(\eta, x, y)dV_g(x) = \tilde{\mathcal{O}}(1).$$

□

F.1.2 Analysis of truncation error

Now we discuss the truncation level needed to guarantee a high accuracy evaluation of heat kernel as required in Proposition 19.

Proposition 22. *Let M be a compact manifold, and assume M is a homogeneous space. With $\frac{1}{\eta} = \tilde{\mathcal{O}}(\log \frac{1}{\varepsilon})$ and $\zeta = \frac{\varepsilon}{\log^2 \frac{1}{\varepsilon}}$, to reach $\|\nu(\eta, x, y) - \nu_l(\eta, x, y)\|_{L^2}^2 = \tilde{\mathcal{O}}(\zeta)$ we need $l = \mathrm{Poly}\,(\log \frac{1}{\varepsilon})$. Consequently, to achieve*

$$\begin{aligned}
&\int_M |\nu(\eta, x, y) - \nu_l(\eta, x, y)|dV_g(x) = \tilde{\mathcal{O}}(\zeta) \qquad \textit{and} \\
&\int_M |\nu(\eta, x, y) - \nu_l(\eta, x, y)|dV_g(y) = \tilde{\mathcal{O}}(\zeta),
\end{aligned}$$

we need $l = \mathrm{Poly}\,(\log \frac{1}{\varepsilon})$.

Proof. [Proof of Proposition 22] Following Azangulov et al. [2022, Proof of Proposition 21] we have

$$\|\nu(\eta, x, y) - \nu_l(\eta, x, y)\|_{L^2}^2 \leq C' l \frac{1}{\eta^2} e^{-\frac{\eta^2 l^{2/d}}{C}}.$$

Take $\frac{1}{\eta} = \log \frac{1}{\varepsilon}$. Recall that in Theorem 8 we require $\zeta = \frac{\varepsilon}{\log^2 \frac{1}{\varepsilon}}$. Requiring $C' l \frac{1}{\eta^2} e^{-\frac{\eta^2 l^{2/d}}{C}} = \tilde{\mathcal{O}}(\zeta)$ is equivalent to

$$C' l \log^2 \frac{1}{\varepsilon} e^{-\frac{\frac{1}{\log^2 \frac{1}{\varepsilon}} l^{2/d}}{C}} \leq \zeta = \frac{\varepsilon}{\log^2 \frac{1}{\varepsilon}}.$$

Take log on both sides, we get $-\frac{\frac{1}{\log^2 \frac{1}{\varepsilon}} l^{2/d}}{C} \leq \log \frac{\varepsilon}{C' l \log^4 \frac{1}{\varepsilon}}$. This further implies

$$l^{2/d} \geq -\log \frac{\varepsilon}{C' l \log^4 \frac{1}{\varepsilon}} C \log^2 \frac{1}{\varepsilon} = (4 \log\log \frac{1}{\varepsilon} + \log \frac{1}{\varepsilon} + \log l + C'') C \log^2 \frac{1}{\varepsilon}.$$

It suffices to take $l = \text{Poly}\,(\log \frac{1}{\varepsilon})$. We verify that $l = \text{Poly}\,(\log \frac{1}{\varepsilon})$ can guarantee the bound:

$$C' l \frac{1}{\eta^2} e^{-\frac{\eta^2 l^{2/d}}{C}} = C' l \log^2 \frac{1}{\varepsilon} e^{-\frac{l^{2/d}}{C \log^2 \frac{1}{\varepsilon}}} = \text{Poly}\,(\log \frac{1}{\varepsilon}) e^{-\text{Poly}\,(\log \frac{1}{\varepsilon})} = \tilde{\mathcal{O}}(\frac{\varepsilon}{\log^2 \frac{1}{\varepsilon}}) = \tilde{\mathcal{O}}(\zeta).$$

On a homogeneous space, both ν and ν_l are stationary [Azangulov et al., 2022]. Hence $\int_M |\nu(\eta, x, y) - \nu_l(\eta, x, y)| dV_g(x)$ does not depend on y, and $\int_M |\nu(\eta, x, y) - \nu_l(\eta, x, y)| dV_g(y)$ does not depend on x. Therefore using Jensen's inequality,

$$\begin{aligned}\int_M |\nu(\eta, x, y) - \nu_l(\eta, x, y)| dV_g(x) &= \tilde{\mathcal{O}}(\|\nu(\eta, x, y) - \nu_l(\eta, x, y)\|_{L_1}) \\ &\leq \tilde{\mathcal{O}}(\|\nu(\eta, x, y) - \nu_l(\eta, x, y)\|_{L_2}).\end{aligned}$$

Note that the same holds for $\int_M |\nu(\eta, x, y) - \nu_l(\eta, x, y)| dV_g(y)$. Hence we get the desired bound, i.e., $\int_M |\nu(\eta, x, y) - \nu_l(\eta, x, y)| dV_g(x) = \tilde{\mathcal{O}}(\zeta)$ and $\int_M |\nu(\eta, x, y) - \nu_l(\eta, x, y)| dV_g(y) = \tilde{\mathcal{O}}(\zeta)$. □

G Proofs for entropy-regularized JKO scheme

Proof. [Proof of Lemma 11] Note that, we have

$$\begin{aligned}H_{\pi^X}(\rho) &= \int_M \rho(x) \log \frac{\rho(x)}{\pi^X} dV_g(x) = \int_M \rho(x) \log \frac{\rho(x)}{\exp(-f(x) - \frac{d(x,y)^2}{2\eta}) \exp(\frac{d(x,y)^2}{2\eta}) C} dV_g(x) \\ &= \int_M \rho(x) (\log \frac{\rho(x)}{C \exp(-f(x) - \frac{d(x,y)^2}{2\eta})} + \log \frac{1}{\exp(\frac{d(x,y)^2}{2\eta})}) dV_g(x) \\ &= \int_M \rho(x) \log \frac{\rho(x)}{C'(y) \tilde{\pi}^{X|Y}(x|y)} dV_g(x) - \int_M \rho(x) \frac{1}{2\eta} d(x,y)^2 dV_g(x) \\ &= H_{\tilde{\pi}^{X|Y=y}}(\rho) - \frac{1}{2\eta} \int_M d(x,y)^2 d\rho + C(y),\end{aligned}$$

where $C = \frac{1}{\int_M e^{-f(x)} dV_g(x)}$, $C'(y)$ and $C(y)$ are some constants that only depends on y. The above computation implies

$$\begin{aligned}\tilde{\pi}^{X|Y=y} &= \arg\min_{\rho \in \mathcal{P}_2(M)} H_{\tilde{\pi}^{X|Y=y}}(\rho) = \arg\min_{\rho \in \mathcal{P}_2(M)} H_{\tilde{\pi}^X}(\rho) + \frac{1}{2\eta} \int_M d(x,y)^2 d\rho + C(y) \\ &= \arg\min_{\rho \in \mathcal{P}_2(M)} H_{\tilde{\pi}^X}(\rho) + \frac{1}{2\eta} W_2^2(\rho, \delta_y) = \text{prox}_{\eta H_{\tilde{\pi}^X}}(\delta_y).\end{aligned}$$

□

Lemma 23. *The minimization problem*

$$\min_{\gamma \in \mathcal{P}_2(M \times M), \gamma^X = \rho^X} \int_{M \times M} \frac{1}{2\eta} d(x,y)^2 \gamma(x,y) dV_g(x) dV_g(y) + H(\gamma),$$

where the constraint means $\int_M \gamma(x,y) dV_g(y) = \rho^X(x)$, has solution of the form

$$\gamma(x,y) \propto \rho^X(x) \tilde{\pi}^{Y|X}(y|x).$$

Proof. [Proof of Lemma 23] Since $\int_M \gamma(x,y) dV_g(y) = \rho^X(x)$, we have

$$\int_M \Big(\int_M \gamma(x,y) dV_g(y) - \rho^X(x) \Big) \beta(x) dV_g(x) = 0, \forall \beta,$$

we can construct the following Lagrangian

$$\int_{M \times M} \frac{1}{2\eta} d(x,y)^2 \gamma(x,y) dV_g(x) dV_g(y) + H(\gamma) - \int_M \Big(\int_M \gamma(x,y) dV_g(y) - \rho^X(x) \Big) \beta(x) dV_g(x).$$

Recall that $H(\gamma) = \int_{M\times M} \gamma \log(\gamma) dV_g(x) dV_g(y)$. We have,

$$\begin{aligned}\lim_{t\to 0} \frac{H(\gamma + t\varphi) - H(\gamma)}{t} &= \lim_{t\to 0} \frac{\int_{M\times M} (\gamma + t\varphi)\log(\gamma + t\varphi) - \gamma\log(\gamma) dV_g(x) dV_g(y)}{t} \\ &= \lim_{t\to 0} \int_{M\times M} \varphi \log(\gamma + t\varphi) + \varphi dV_g(x) dV_g(y) \\ &= \int_{M\times M} \varphi(\log(\gamma) + 1) dV_g(x) dV_g(y).\end{aligned}$$

For any function f, denote $I_f(\gamma) = \int_{M\times M} \gamma(x,y) f(x,y) dV_g(x) dV_g(y)$. We then have

$$\begin{aligned}\lim_{t\to 0} \frac{I_f(\gamma + t\varphi) - I_f(\gamma)}{t} &= \lim_{t\to 0} \frac{\int_{M\times M} (\gamma + \varphi t) f - \gamma f dV_g(x) dV_g(y)}{t} \\ &= \int_{M\times M} \varphi f dV_g(x) dV_g(y).\end{aligned}$$

Thus the variation of Lagrangian is given by

$$\int_{M\times M} \varphi \cdot \Big(\frac{1}{2\eta} d(x,y)^2 + \log(\gamma) + 1 - \beta(x)\Big) dV_g(x) dV_g(y).$$

We want the above to be zero for all φ. Thus we need $\frac{1}{2\eta} d(x,y)^2 + \log(\gamma) + 1 - \beta(x) = 0$ which is equivalent to

$$\gamma(x,y) = e^{\beta(x) - \frac{1}{2\eta} d(x,y)^2 - 1}.$$

This implies $\gamma(x,y) \propto e^{\beta(x) - \frac{1}{2\eta} d(x,y)^2}$ Integrating with respect to the y variable, we get

$$\rho^X(x) = \int_M \gamma(x,y) dV_g(y) = \int_M e^{\beta(x) - \frac{1}{2\eta} d(x,y)^2 - 1} dV_g(y) \propto e^{\beta(x)} \int_M e^{-\frac{1}{2\eta} d(x,y)^2} dV_g(y).$$

It then follows that

$$\gamma(x,y) \propto \rho^X(x) \frac{e^{-\frac{1}{2\eta} d(x,y)^2}}{\int_M e^{-\frac{1}{2\eta} d(x,y)^2} dV_g(y)} = \rho^X(x) \tilde{\pi}^{Y|X}(y|x).$$

□

Lemma 24. *The minimization problem*

$$\min_{\gamma \in \mathcal{P}_2(M\times M), \gamma^Y = \rho^Y} \int_{M\times M} (f(x) + \frac{1}{2\eta} d(x,y)^2) \gamma(x,y) dV_g(x) dV_g(y) + H(\gamma),$$

where the constraint means $\int_M \gamma(x,y) dV_g(x) = \rho^Y(y)$, has solution of the form

$$\gamma(x,y) \propto \rho^Y(y) \tilde{\pi}^{X|Y}(x|y).$$

Proof. [Proof of Lemma 24] The proof follows similarly to that of Lemma 23. Since $\int_M \gamma(x,y) dV_g(x) = \rho^Y(y)$, we have

$$\int_M \Big(\int_M \gamma(x,y) dV_g(x) - \rho^Y(y)\Big) \beta(y) dV_g(y) = 0, \forall \beta.$$

We first constructing the following Lagrangian:

$$\int_{M\times M} (f(x) + \frac{1}{2\eta} d(x,y)^2) \gamma(x,y) dV_g(x) dV_g(y) + H(\gamma) - \int_M \Big(\int_M \gamma(x,y) dV_g(x) - \rho^Y(y)\Big) \beta(y) dV_g(y).$$

Recall that $H(\gamma) = \int_{M\times M} \gamma \log(\gamma) dV_g(x) dV_g(y)$. Then, we have

$$\lim_{t\to 0} \frac{H(\gamma + t\varphi) - H(\gamma)}{t} = \lim_{t\to 0} \frac{\int_{M\times M} (\gamma + t\varphi)\log(\gamma + t\varphi) - \gamma\log(\gamma) dV_g(x) dV_g(y)}{t}$$

$$= \lim_{t\to 0} \int_{M\times M} \varphi \log(\gamma + t\varphi) + \varphi dV_g(x) dV_g(y)$$
$$= \int_{M\times M} \varphi(\log(\gamma) + 1) dV_g(x) dV_g(y).$$

For any function f, denote $I_f(\gamma) = \int_{M\times M} \gamma(x,y) f(x,y) dV_g(x) dV_g(y)$. We have

$$\lim_{t\to 0} \frac{I_f(\gamma + t\varphi) - I_f(\gamma)}{t} = \lim_{t\to 0} \frac{\int_{M\times M} (\gamma + \varphi t) f - \gamma f dV_g(x) dV_g(y)}{t}$$
$$= \int_{M\times M} \varphi f dV_g(x) dV_g(y).$$

Thus the variation of Lagrangian is

$$\int_{M\times M} \varphi \cdot \Big(f(x) + \frac{1}{2\eta} d(x,y)^2 + \log(\gamma) + 1 - \beta(y) \Big) dV_g(x) dV_g(y).$$

We want the above to be zero for all φ. Thus we need $f(x) + \frac{1}{2\eta} d(x,y)^2 + \log(\gamma) + 1 - \beta(y) = 0$ which is equivalent to

$$\gamma(x,y) = e^{-f(x) + \beta(y) - \frac{1}{2\eta} d(x,y)^2 - 1}.$$

This implies $\gamma(x,y) \propto e^{\beta(y) - f(x) - \frac{1}{2\eta} d(x,y)^2}$. Hence we can integrate with respect to the x variable and get

$$\rho^Y(y) = \int_M e^{-f(x) + \beta(y) - \frac{1}{2\eta} d(x,y)^2 - 1} dV_g(x) \propto e^{\beta(y)} \int_M e^{-f(x) - \frac{1}{2\eta} d(x,y)^2} dV_g(x).$$

Therefore, we obtain

$$\gamma(x,y) \propto \rho^Y(y) \frac{e^{-f(x) - \frac{1}{2\eta} d(x,y)^2}}{\int_M e^{-f(x) - \frac{1}{2\eta} d(x,y)^2} dV_g(x)} = \rho^Y(y) \tilde{\pi}^{X|Y}(x|y).$$

□

Proof. [Proof of Theorem 12] By definition we have

$$\arg\min_{\chi \in \mathcal{P}_2(M)} \frac{1}{2\eta} W_{2,2\eta}^2(\tilde{\rho}_k^X, \chi)$$
$$= \arg\min_{\chi \in \mathcal{P}_2(M): \gamma \in C(\tilde{\rho}_k^X, \chi)} \int_{M\times M} \frac{1}{2\eta} d(v,w)^2 \gamma(v,w) dV_g(v) dV_g(w) + H(\gamma).$$

By Lemma 24, we know the solution of

$$\min_{\gamma \in \mathcal{P}_2(M\times M), \gamma^X = \tilde{\rho}_k^X} \int_{M\times M} \frac{1}{2\eta} d(x,y)^2 \gamma(x,y) dV_g(x) dV_g(y) + H(\gamma)$$

is $\gamma(x,y) \propto \tilde{\rho}_k^X(x) e^{-\frac{1}{2\eta} d(x,y)^2}$. Hence the Y-marginal of inexact proximal sampler satisfies

$$\chi(y) = \int_M \tilde{\rho}_k^X(x) e^{-\frac{1}{2\eta} d(x,y)^2} dV_g(x) = \int_M \tilde{\rho}_k^X(x) \tilde{\pi}^{Y|X}(y|x) dV_g(x) = \tilde{\rho}_k(y).$$

Similarly,

$$\arg\min_{\chi \in \mathcal{P}_2(M)} \frac{1}{2\eta} W_{2,2\eta}^2(\tilde{\rho}_k^Y, \chi) + \int f d\chi$$
$$= \arg\min_{\chi \in \mathcal{P}_2(M): \gamma \in C(\tilde{\rho}_k^X, \chi)} \int_{M\times M} (f(x) + \frac{1}{2\eta} d(v,w)^2) \gamma(v,w) dV_g(v) dV_g(w) + H(\gamma),$$

and its solution is $\chi(x) = \int_M \tilde{\rho}^Y(y) \tilde{\pi}^{X|Y}(x|y) dV_g(y) = \tilde{\rho}_{k+1}^X(x)$. □

H Auxiliary results

H.1 Diffusion process on manifold

It is well known that the law of the following SDE $dX_t = -b(X_t)dt + dB_t$ is related to the Fokker-Planck equation $\partial_t \rho_t = \operatorname{div}(\rho_t b(X_t) + \frac{1}{2}\operatorname{grad}\ \rho_t)$. Here we provide a proof for completeness.

Lemma 25. *Let B_t denote Brownian motion on a Riemannian manifold M. For SDE $dX_t = -b(X_t)dt + dB_t$, the corresponding Fokker-Planck equation is*

$$\partial_t \rho_t = \operatorname{div}(\rho_t b(X_t) + \frac{1}{2}\operatorname{grad}\ \rho_t).$$

Proof. The infinitesimal generator of the SDE is $Lf = -\langle \operatorname{grad}\ f, b\rangle + \frac{1}{2}\Delta f$ Cheng et al. [2022]. We compute the adjoint of L which is defined by $\int_M fL^* h dV_g = \int_M hLf dV_g$. By divergence theorem, we have

$$\begin{aligned}\frac{1}{2}\int_M \operatorname{div}(\operatorname{grad}\ f)h dV_g &= -\frac{1}{2}\int_M \langle \operatorname{grad}\ h, \operatorname{grad}\ f\rangle dV_g = \frac{1}{2}\int_M \operatorname{div}(\operatorname{grad}\ h) f dV_g \\ -\int_M \langle \operatorname{grad}\ f, bh\rangle dV_g &= \int_M \operatorname{div}(bh) f dV_g.\end{aligned}$$

Hence

$$\begin{aligned}\int_M hLf dV_g &= \int_M -h\langle \operatorname{grad}\ f, b\rangle + \frac{1}{2}h\Delta f dV_g \\ &= \int_M f \operatorname{div}(bh) + \frac{1}{2}f\Delta h dV_g = \int_M f(\operatorname{div}(bh) + \frac{1}{2}\Delta h) dV_g.\end{aligned}$$

Thus we obtained $L^* h = \operatorname{div}(bh) + \frac{1}{2}\Delta h$. By Kolmogorov forward equation [Bakry et al., 2014, Equation 1.5.2], we get

$$\partial_t \rho_t = L^* \rho_t = \operatorname{div}(b(X_t)\rho_t) + \frac{1}{2}\Delta \rho_t = \operatorname{div}(b(X_t)\rho_t + \frac{1}{2}\operatorname{grad}\ \rho_t).$$

□

We briefly mention some properties of Markov semigroup. The following results are from Bakry et al. [2014, Section 1.2].

Definition 26.

1. *Given a markov process, the assoicated markov semigroup $(P_t)_{t\geq 0}$ is defined as (for suitable f)*

$$P_t f(x) = \mathbb{E}[f(X_t)|X_0 = x], \forall t \geq 0.$$

2. *Let ρ be the law of X_0, then $P_t^* \rho$ is the law of X_t. We have*

$$\int_M P_t f d\rho = \int_M f d(P_t^* \rho).$$

3. *Markov operators $(P_t)_{t\geq 0}$ can be represented by kernels corresponding to the transition probabilities of the associated Markov process:*

$$P_t f(x) = \int_M f(y) p_t(x, y) dV_g(y), \forall t \geq 0, x \in M.$$

Thus by definition, we have

$$\mathbb{E}[f(X_t)|X_0 = x] = P_t f(x) = \int_M f(y) p_t(x, y) dV_g(y).$$

H.2 Log-Sobolev inequality and heat flow

In the sampling literature, the log-Sobolev inequality is usually written in the following form:

$$\int_M f^2 \log f^2 d\nu - \int_M f^2 d\nu \log \int_M f^2 d\nu \le \frac{2}{\alpha} \int_M \| \operatorname{grad} f\|^2 d\nu, \forall f$$
$$H_\nu(\rho) \le \frac{1}{2\alpha} J_\nu(\rho), \forall \rho.$$

In the Euclidean setting, we know if μ_1, μ_2 satisfy α_1, α_2-LSI respectively, then their convolution $\mu_1 * \mu_2$ satisfies LSI with constant $\frac{1}{\frac{1}{\alpha_1}+\frac{1}{\alpha_2}}$, see Chewi [2023, Proposition 2.3.7]. In particular, if we take one of μ to be $\nu(t, x, y)$ (which is a Gaussian in the Euclidean setting), since the Gaussian density satisfies LSI, we have the following result.

Fact 27. *Consider Euclidean space. Let μ be a probability measure that satisfies α-LSI. Then its propogation along heat flow, denoted by $\mu_t = \mu * \nu_t$, also satisfies LSI with constant $\frac{1}{\frac{1}{\alpha}+t} = \frac{\alpha}{1+t\alpha}$. Here ν_t denote the probability measure corresponding to heat flow for time t.*

On a Riemannian manifold, the density for Brownian motion satisfies LSI.

Theorem 28. *[Hsu, 1997, Theorem 3.1] Suppose M is a complete, connected manifold with $\operatorname{Ric}_M \ge -c$. Here $c \ge 0$. Then for any smooth function on M, we have*

$$\int_M f^2 \log |f| d\nu_{o,s} \le \frac{e^{cs}-1}{c} \| \operatorname{grad} f\|^2_{\nu_{o,s}} + \|f\|^2_{\nu_{o,s}} \log \|f\|_{\nu_{o,s}}.$$

With $\kappa = -c$, we know the Brownian motion density for time t satisfies LSI with constant $\alpha = \frac{\kappa}{1-e^{-\kappa t}}$.

As a special case $M = \mathbb{R}^d$, we have $c = 0$. Hence, the LSI constant became $\lim_{c\to 0} \frac{e^{ct}-1}{c} = t$. That is, (with ν representing the measure for Brownian motion with time t) $H_\nu(\rho) \le \frac{t}{2} I_\nu(\rho), \forall \rho$. So the LSI constant for Brownian motion is $\alpha_\nu = \frac{1}{t}$.

In the following, we prove that on a Riemannian manifold, such a fact is still true. We follow the idea by Collet and Malrieu [2008, Theorem 4.1]. For notations, we denote $\Gamma(f) = \Gamma(f, f) = \| \operatorname{grad} f\|^2_g$. We also require the following intermediate result.

Lemma 29 (Theorem 5.5.2 in Bakry et al. [2014])**.** *For Markov triple with semigroup $(P_t)_{t\ge 0}$, the followings are equivalent:*

1. $\sqrt{\Gamma(P_s f)} \le e^{-\beta s} P_s \sqrt{\Gamma(f)}$.
2. $P_s(f \log f) - P_s f \log(P_s f) \le c(s) P_s(\frac{\Gamma(f)}{f})$ *where* $c(s) = \frac{1-e^{-2\beta s}}{2\beta}$.

Corollary 30. *With P_t denote manifold Brownian motion, we have*

1. $\sqrt{\Gamma(P_t f)} \le e^{-\frac{\kappa}{2} t} P_t \sqrt{\Gamma(f)}$.
2. $P_t(f \log f) - P_t f \log(P_t f) \le c(t) P_t(\frac{\Gamma(f)}{f})$ *where* $c(t) = \frac{1-e^{-\kappa t}}{2\kappa}$.

Proof. [Proof of Corollary 30] For the second item, we can replace f by g^2 for some g.

$$\int_M g^2 \log g^2 d\nu_s - \int_M g^2 d\nu_s \log(\int_M g^2 d\nu_s) \le \frac{1-e^{-2\beta s}}{2\beta} \int_M \frac{(2g)^2 \| \operatorname{grad} g\|^2}{g^2} d\nu_s$$
$$= 4\frac{1-e^{-2\beta s}}{2\beta} \int_M \| \operatorname{grad} g\|^2 d\nu_s.$$

Now we already know the manifold Brownian motion density ν_t satisfies $\frac{\kappa}{1-e^{-\kappa t}}$-LSI, i.e.,

$$\int_M f^2 \log f^2 d\nu_t - \int_M f^2 d\nu_t \log \int_M f^2 d\nu_t \le 2\frac{1-e^{-\kappa t}}{\kappa} \int_M \| \operatorname{grad} f\|^2 d\nu_t, \forall f.$$

So we know, with P_t representing manifold Brownian motion, $P_t(f\log f) - P_t f\log(P_t f) \leq c(t)P_t(\frac{\Gamma(f)}{f})$, where

$$2\frac{1-e^{-\kappa t}}{\kappa} = 4\frac{1-e^{-2\beta s}}{2\beta}.$$

Hence we know β can be taken as κ, s corresponds to $\frac{1}{2}t$. So we get $\sqrt{\Gamma(P_t f)} \leq e^{-\frac{\kappa}{2}t}P_t\sqrt{\Gamma(f)}$. $\square$

Proposition 31. *Let M be a Riemannian manifold with Ricci curvature bounded below by κ. Let ρ_0 be any initial distribution. Assume ρ_0 satisfies LSI with constant $\frac{1}{d_0}$:*

$$\int_M g^2\log g^2 d\rho_0 - \int_M g^2 d\rho_0 \log\int_M g^2 d\rho_0 \leq 2d_0\int_M \|\text{ grad }g\|^2 d\rho_0, \forall g.$$

Then the propagation of ρ_0 along heat flow, denoted as ρ_t, satisfies LSI with constant

$$\frac{1}{2c(t)+d_0e^{-\kappa t}} = \frac{\kappa}{1-e^{-\kappa t}+\kappa d_0 e^{-\kappa t}} = \frac{\kappa e^{\kappa t}}{e^{\kappa t}-1+\kappa d_0},$$

where $c(t) = \frac{1-e^{-\kappa t}}{2\kappa}$. If $\kappa \geq 0$, we have $c(t) \leq \frac{1}{2}t$.

$$\frac{1}{2c(t)+d_0e^{-\kappa t}} \geq \frac{1}{t+d_0e^{-\kappa t}} \geq \frac{1}{t+d_0}.$$

Proof. [Proof of Proposition 31] Since ρ_0 satisfies LSI with constant $\frac{1}{d_0}$, equivalently with f replace g^2, we get

$$\int_M f\log f d\rho_0 - \int_M f d\rho_0 \log\int_M f d\rho_0 \leq 2d_0\int_M \|\text{ grad }\sqrt{f}\|^2 d\rho_0 = \frac{d_0}{2}\int_M \frac{\|\text{ grad }f\|^2}{f}d\rho_0.$$

For $g > 0$, using Corollary 30, we know the manifold Brownian motion (here represented by P_t) satisfies

$$P_t(g\log g) - P_t g\log(P_t g) \leq c(t)P_t(\frac{\Gamma(g)}{g}),$$

where $c(t) = \frac{1-e^{-\kappa t}}{2\kappa}$. Using property of markov semigroup, we have

$$\int_M g\log g d\rho_t = \int_M P_t(g\log g)d\rho_0 \leq c(t)\int_M P_t(\frac{\Gamma(g)}{g})d\rho_0 + \int_M P_t g\log(P_t g)d\rho_0.$$

Hence

$$\begin{aligned}
&\int_M g\log g d\rho_t - \int_M g d\rho_t \log\int_M g d\rho_t\\
\leq& c(t)\int_M P_t(\frac{\Gamma(g)}{g})d\rho_0 + \int_M P_t g\log(P_t g)d\rho_0 - \int_M P_t(g)d\rho_0\log\int_M P_t(g)d\rho_0\\
\leq& c(t)\int_M P_t(\frac{\Gamma(g)}{g})d\rho_0 + \frac{d_0}{2}\int_M \frac{\Gamma(P_t g)}{P_t g}d\rho_0\\
\leq& c(t)\int_M P_t(\frac{\Gamma(g)}{g})d\rho_0 + \frac{d_0}{2}e^{-\kappa t}\int_M \frac{(P_t\sqrt{\Gamma(g)})^2}{P_t g}d\rho_0\\
\leq& c(t)\int_M P_t(\frac{\Gamma(g)}{g})d\rho_0 + \frac{d_0}{2}e^{-\kappa t}\int_M P_t(\frac{\Gamma(g)}{g})d\rho_0\\
=&\Big(c(t)+\frac{d_0}{2}e^{-\kappa t}\Big)\int_M P_t(\frac{\Gamma(g)}{g})d\rho_0 = \Big(c(t)+\frac{d_0}{2}e^{-\kappa t}\Big)\int_M \frac{\Gamma(g)}{g}d\rho_t,
\end{aligned}$$

where the third inequality is due to Corollary 30, and in the last inequality we used Cauchy-Schwarz inequality:

$$P_t(\frac{\Gamma(f)}{f})P_t(f) = \mathbb{E}[\frac{\|\text{ grad }f\|^2}{f}]\mathbb{E}[f] \geq \mathbb{E}\Big[\sqrt{\frac{\|\text{ grad }f\|^2}{f}}f\Big]^2 = \mathbb{E}[\|\text{ grad }f\|]^2$$

$$= (P_t\sqrt{\Gamma(f)})^2.$$

Hence we know ρ_t satisfies LSI with constant

$$\frac{1}{2c(t)+d_0e^{-\kappa t}} = \frac{\kappa}{1-e^{-\kappa t}+\kappa d_0 e^{-\kappa t}}.$$

□

Note that we have $\lim_{\kappa\to 0}\frac{1}{2c(t)+d_0e^{-\kappa t}} = \frac{1}{t+d_0}$. This means that, we can recover the result for Euclidean space in the limit.

H.3 Extension of Proposition 31 to Phi-Sobolev inequality

In the last subsection, we showed in Proposition 31 that on a Riemannian manifold, LSI is preserved along the propagation of heat flow. In this subsection, we extend the result to the setting of Phi-Sobolev inequality. We will prove the following proposition. This result will be useful for proving convergence of the Riemannian proximal sampler under χ^2 divergence and Poincaré inequality (Theorem 6 item 2, the proof is in Appendix D.4).

Proposition 32. *Let M be a Riemannian manifold with Ricci curvature bounded below by κ. Assume $\phi : I \to \mathbb{R}$ is a C^4 function and convex, where I is an interval of $\mathbb{R}$. Further assume $\frac{1}{\phi''}$ is concave on I. Let ρ_0 be any initial distribution. Assume ρ_0 satisfies ϕ-entropy inequality [Bakry et al., 2014, Section 7.6.1] with constant $\frac{1}{d_0}$:*

$$\int_M \phi(g)d\rho_0 - \phi(\int_M g d\rho_0) \le \frac{d_0}{2}\int_M \phi''(g)\Gamma(g)d\rho_0, \forall g.$$

Then the propagation of ρ_0 along heat flow, denoted as ρ_t, satisfies ϕ-entropy inequality with constant

$$\frac{1}{2c(t)+d_0e^{-\kappa t}} = \frac{\kappa}{1-e^{-\kappa t}+\kappa d_0 e^{-\kappa t}} = \frac{\kappa e^{\kappa t}}{e^{\kappa t}-1+\kappa d_0},$$

where $c(t) = \frac{1-e^{-\kappa t}}{2\kappa}$.

We first present some useful lemmas.

Lemma 33. *Let ϕ be such that $-\frac{1}{\phi''(x)}$ is convex. Then*

$$P_t(\Gamma(g)\phi''(g))\frac{1}{\phi''(P_tg)} \ge (P_t(\sqrt{\Gamma(g)}))^2$$

In particular, when $\phi(x) = (x-1)^2$, we know $-\frac{1}{\phi''(x)} = -\frac{1}{2}$ is a constant, and therefore $P_t(\Gamma(g)) \ge (P_t(\sqrt{\Gamma(g)}))^2$.

Proof. Using Jensen's inequality, since $\frac{1}{\phi''(x)}$ is concave, we have $\frac{1}{\phi''(\mathbb{E}[g])} \ge \mathbb{E}[\frac{1}{\phi''(g)}]$.

$$P_t(\Gamma(g)\phi''(g))\frac{1}{\phi''(g)} = \mathbb{E}[\Gamma(g)\phi''(g)]\frac{1}{\phi''(\mathbb{E}[g])}$$
$$\ge \mathbb{E}[\Gamma(g)\phi''(g)]\mathbb{E}[\frac{1}{\phi''(g)}] \ge \mathbb{E}[\sqrt{\Gamma(g)}]^2 = (P_t(\sqrt{\Gamma(g)}))^2$$

where in the last inequality we used Cauchy-Schwarz. □

Lemma 34. *For diffusion Markov triple with semigroup $(P_t)_{t\ge 0}$, the followings are equivalent:*

1. *The curvature dimension $CD(\beta,\infty)$ holds for some $\beta \in \mathbb{R}$.*
2. *$\sqrt{\Gamma(P_sf)} \le e^{-\beta s}P_s\sqrt{\Gamma(f)}$.*
3. *$P_s(f\log f) - P_sf\log(P_sf) \le c(s)P_s(\frac{\Gamma(f)}{f})$.*
4. *$P_s(f^2) - (P_s(f))^2 \le 2c(s)P_s(\Gamma(f))$.*

where $c(s) = \frac{1-e^{-2\beta s}}{2\beta}$. *Furthermore, the first item implies* $P_t(\phi(g)) - \phi(P_t g) \leq c(t) P_t(\phi''(g)\Gamma(g))$

Proof. See Theorem 4.7.2 and Theorem 5.5.2 in Bakry et al. [2014], and Theorem 2.1 in Chafaï [2004]. □

Corollary 35. *With* P_t *denote manifold Brownian motion, we have*

1. $\sqrt{\Gamma(P_t f)} \leq e^{-\frac{\kappa}{2}t} P_t \sqrt{\Gamma(f)}$.
2. $P_t(f \log f) - P_t f \log(P_t f) \leq c(t) P_t(\frac{\Gamma(f)}{f})$
3. $P_s(f^2) - (P_s(f))^2 \leq 2c(t) P_s(\Gamma(f))$.
4. $P_t(\phi(f)) - \phi(P_t f) \leq c(t) P_t(\phi''(f)\Gamma(f))$

where $c(t) = \frac{1-e^{-\kappa t}}{2\kappa}$.

Proof. [Proof of Corollary 35] For the second item, we can replace f by g^2 for some g.

$$\begin{aligned}\int_M g^2 \log g^2 d\nu_s - \int_M g^2 d\nu_s \log(\int_M g^2 d\nu_s) &\leq \frac{1-e^{-2\beta s}}{2\beta} \int_M \frac{(2g)^2 \| \operatorname{grad} g\|^2}{g^2} d\nu_s \\ &= 4\frac{1-e^{-2\beta s}}{2\beta} \int_M \| \operatorname{grad} g\|^2 d\nu_s.\end{aligned}$$

Now we already know the manifold Brownian motion density ν_t satisfies $\frac{\kappa}{1-e^{-\kappa t}}$-LSI, i.e.,

$$\int_M f^2 \log f^2 d\nu_t - \int_M f^2 d\nu_t \log \int_M f^2 d\nu_t \leq 2\frac{1-e^{-\kappa t}}{\kappa} \int_M \| \operatorname{grad} f\|^2 d\nu_t, \forall f.$$

So we know, with P_t representing manifold Brownian motion, $P_t(f \log f) - P_t f \log(P_t f) \leq c(t) P_t(\frac{\Gamma(f)}{f})$, where

$$2\frac{1-e^{-\kappa t}}{\kappa} = 4\frac{1-e^{-2\beta s}}{2\beta}.$$

Hence we know β can be taken as κ, s corresponds to $\frac{1}{2}t$. So we get item 1, 3 and 4. □

Now we are ready to prove Proposition 32

Proof. [Proof of Proposition 32]

For $g > 0$, using Corollary 35, we know the manifold Brownian motion (here represented by P_t) satisfies

$$P_t(\phi(g)) - \phi(P_t g) \leq c(t) P_t(\phi''(g)\Gamma(g)),$$

where $c(t) = \frac{1-e^{-\kappa t}}{2\kappa}$. Using property of markov semigroup, we have

$$\int_M \phi(g) d\rho_t = \int_M P_t(\phi(g)) d\rho_0 \leq c(t) \int_M P_t(\phi''(g)\Gamma(g)) d\rho_0 + \int_M \phi(P_t g) d\rho_0.$$

Hence

$$\begin{aligned}&\int_M \phi(g) d\rho_t - \phi(\int_M g d\rho_t) \\ \leq& c(t) \int_M P_t(\phi''(g)\Gamma(g)) d\rho_0 + \int_M \phi(P_t g) d\rho_0 - \phi(\int_M P_t(g) d\rho_0) \\ \leq& c(t) \int_M P_t(\phi''(g)\Gamma(g)) d\rho_0 + \frac{d_0}{2} \int_M \phi''(P_t g)\Gamma(P_t g) d\rho_0 \\ \leq& c(t) \int_M P_t(\phi''(g)\Gamma(g)) d\rho_0 + \frac{d_0}{2} e^{-\kappa t} \int_M \phi''(P_t g)(P_t \sqrt{\Gamma(g)})^2 d\rho_0\end{aligned}$$

$$\leq c(t)\int_M P_t(\phi''(g)\Gamma(g))d\rho_0 + \frac{d_0}{2}e^{-\kappa t}\int_M P_t(\Gamma(g)\phi''(g))d\rho_0$$
$$=\Big(c(t)+\frac{d_0}{2}e^{-\kappa t}\Big)\int_M P_t(\phi''(g)\Gamma(g))d\rho_0 = \Big(c(t)+\frac{d_0}{2}e^{-\kappa t}\Big)\int_M \phi''(g)\Gamma(g)d\rho_t,$$

where the third inequality is due to Corollary 30, and in the last inequality we used Lemma 33.

Hence we know ρ_t satisfies ϕ-entropy inequality with constant

$$\frac{1}{2c(t)+d_0e^{-\kappa t}} = \frac{\kappa}{1-e^{-\kappa t}+\kappa d_0 e^{-\kappa t}}.$$

□

We remark that the case $\phi(x) = (x-1)^2$ recovers Poincaré inequality, and $\phi(x) = x\log x$ recovers log-Sobolev inequality.

H.4 Total variation distance

Lemma 36. *For TV distance, we have*

$$\|\rho^{(1)}-\rho^{(2)}\|_{TV} := \sup_{A\subseteq M}|\rho^{(1)}(A)-\rho^{(2)}(A)| = \frac{1}{2}\int_M |\frac{d\rho^{(1)}}{dV_g}-\frac{d\rho^{(2)}}{dV_g}|dV_g,$$

and

$$\|\rho^{(1)}-\rho^{(2)}\|_{TV} = \frac{1}{2}\sup_{f:M\to[-1,1]}|\int f d\rho^{(1)} - \int f d\rho^{(2)}|.$$

Proof. [Proof of Lemma 36] Denote the set at which supremum is achieved to be $A_* = \{x\in M : \rho^{(2)}(x)\geq\rho^{(1)}(x)\}$. Denote $\rho^{(2)},\rho^{(1)}$ to be the measure, or corresponding probability density function with respect to the Riemannian volumn form, when appropriate.

$$\int_M |\rho^{(2)}(x)-\rho^{(1)}(x)|dV_g(x) = \int_A |\rho^{(2)}(x)-\rho^{(1)}(x)|dV_g(x) + \int_{A^c}|\rho^{(2)}(x)-\rho^{(1)}(x)|dV_g(x)$$
$$=\int_{A_*}|\rho^{(2)}(x)-\rho^{(1)}(x)|dV_g(x) + \int_{A_*^c}|\rho^{(2)}(x)-\rho^{(1)}(x)|dV_g(x)$$
$$=\int_{A_*}\rho^{(2)}(x)-\rho^{(1)}(x)dV_g(x) + \int_{A_*^c}\rho^{(1)}(x)-\rho^{(2)}(x)dV_g(x)$$
$$=\rho^{(2)}(A_*)-\rho^{(2)}(A_*^c)-\rho^{(1)}(A_*)+\rho^{(1)}(A_*^c) = 2\rho^{(2)}(A_*)-1-2\rho^{(1)}(A_*)+1$$
$$=2(\rho^{(2)}(A_*)-\rho^{(1)}(A_*)) = 2\|\rho^{(2)}-\rho^{(1)}\|_{TV}.$$

Now we prove the second equation.

$$\Big|\int_M f d\rho^{(2)} - \int_M f d\rho^{(1)}\Big| \leq \int_M \Big|f(x)(\rho^{(2)}(x)-\rho^{(1)}(x))\Big|dV_g(x)$$
$$\leq \sup_{x\in M}|f(x)|\int_M|\rho^{(2)}(x)-\rho^{(1)}(x)|dV_g(x)$$
$$=\int_M|\rho^{(2)}(x)-\rho^{(1)}(x)|dV_g(x) = 2\|\rho^{(2)}-\rho^{(1)}\|_{TV}.$$

When $f = 1_{A_*}-1_{A_*^c}$,

$$\Big|\int_M f d\rho^{(2)} - \int_M f d\rho^{(1)}\Big| = \Big|\rho^{(2)}(A_*)-\rho^{(2)}(A_*^c)-\rho^{(1)}(A_*)+\rho^{(1)}(A_*^c)\Big|$$
$$= 2\rho^{(2)}(A_*)-2\rho^{(1)}(A_*) = 2\|\rho^{(2)}-\rho^{(1)}\|_{TV}.$$

□

Lemma 37. *Let $\rho^{(1)}, \rho^{(2)}$ be probability measures. Let $\rho_t^{(1)}, \rho_t^{(2)}$ denote propagation of $\rho^{(1)}, \rho^{(2)}$ along heat flow on M, with $\rho_0^{(1)} = \rho^{(1)}$, $\rho_0^{(2)} = \rho^{(2)}$. We have*

$$\|\rho_t^{(1)} - \rho_t^{(2)}\|_{TV} \leq \|\rho^{(1)} - \rho^{(2)}\|_{TV}.$$

Proof. [Proof of Lemma 37] By definition we have that for all f,

$$\mathbb{E}[f(X_t)|X_0 = x] = P_t f(x) = \int_M f(y) p_t(x, y) dV_g(y).$$

Assuming $X_0 \sim \rho^{(1)}$, we get

$$\begin{aligned}\int_M f(x)\rho_t^{(1)}(x)dV_g(x) = \mathbb{E}[f(X_t)] &= \int P_t f(x) d\rho^{(1)}(x) \\ &= \int_M \int_M f(y) p_t(x, y) dV_g(y) \rho^{(1)}(x) dV_g(x) \\ &= \int_M g(x)\rho^{(1)}(x)dV_g(x).\end{aligned}$$

Where we denote $\int_M f(y)p_t(x, y)dV_g(y) = g(x)$. Note that

$$|g(x)| \leq \int_M |f(y)| p_t(x, y) dV_g(y) \leq \int_M p_t(x, y) dV_g(y) = 1.$$

Hence

$$\begin{aligned}\|\rho_t^{(1)} - \rho_t^{(2)}\|_{TV} =& \frac{1}{2} \sup_{f:M\to[-1,1]} |\int f d\rho_t^{(1)} - \int f d\rho_t^{(2)}| \\ =& \frac{1}{2} \sup_{f:M\to[-1,1]} |\int_M \int_M f(y) p_t(x, y) dV_g(y) \rho^{(1)}(x) dV_g(x) \\ & - \int_M \int_M f(y) p_t(x, y) dV_g(y) \rho^{(2)}(x) dV_g(x)| \\ =& \frac{1}{2} \sup_{f:M\to[-1,1]} |\int_M g(x)\rho^{(1)}(x)dV_g(x) - \int_M g(x)\rho^{(2)}(x)dV_g(x)| \\ \leq& \frac{1}{2} \sup_{g:M\to[-1,1]} |\int_M g(x)\rho^{(1)}(x)dV_g(x) - \int_M g(x)\rho^{(2)}(x)dV_g(x)| \\ =& \|\rho^{(1)} - \rho^{(2)}\|_{TV}.\end{aligned}$$

□

I Concrete example: hypersphere

I.1 Truncation method on hypersphere

Let M be a hypersphere. Previously, we discussed some existing results which provided a bound on the L_2 norm of $\nu_l - \nu$. For hypersphere, we can derive a bound in L_∞ norm, see subsection I.1.2. We also consider the situation that the acceptance rate in rejection sampling might exceed 1, and show that for such a situation, rejection sampling can still produce a high-accuracy sample.

Let $V_{\mathsf{MBI}}(y), V_{\mathsf{RHK}}(x)$ denote the acceptance rate in rejection sampling. Recall that $\hat{\pi}^{Y|X} \propto \nu_l(\eta, x, y)$ and $\hat{\pi}^{X|Y} \propto e^{-f(x)}\nu_l(\eta, x, y)$. In the actual rejection sampling implementation, if for example in Brownian motion implementation, it happens that there exists $y \in M$, s.t. $V(y) > 1$, then the output for rejection sampling will not be equal to $\hat{\pi}^{Y|X}$. For such situations, denote $\overline{V}_{\mathsf{MBI}}(y) = \min\{1, V_{\mathsf{MBI}}(y)\}$ and $\overline{V}_{\mathsf{RHK}}(x) = \min\{1, V_{\mathsf{RHK}}(x)\}$. Note that $\overline{V}_{\mathsf{MBI}}(y)$ and $\overline{V}_{\mathsf{RHK}}(x)$ are the actual acceptance rate in rejection sampling. we denote the corresponding rejection sampling output by $\overline{\pi}^{Y|X}, \overline{\pi}^{X|Y}$, respectively.

Intuitively, the region $B_x(r)$ near x carries most of the probability for both Riemannian Gaussian distribution $\mu(t,x,y)$ as well as Brownian motion $\nu(t,x,y)$, when the variable t is suitably small. Thus instead of choosing parameter to guarantee $V_{\mathsf{RHK}}(x), V_{\mathsf{MBI}}(y) \le 1, \forall x,y \in M$, it suffices to guarantee $V_{\mathsf{RHK}}(x) \le 1, \forall x \in B_y(r)$ and $V_{\mathsf{MBI}}(y) \le 1, \forall y \in B_x(r)$ for some r.

Let L_1 be the Lipschitz constant of f. In the rejection sampling algorithm, we will use $\frac{1}{\eta} = L_1^2 d \log \frac{1}{\zeta}$ (i.e., $\eta = \frac{1}{L_1^2 d \log \frac{1}{\zeta}}$), $\exp(-\frac{d(x,y)^2}{2(s\eta)}) = \exp(-\frac{d(x,y)^2}{2(\frac{1}{L_1^2(d-2)\log\frac{1}{\zeta}})})$ as proposal Riemannian Gaussian distribution.

Define

$$V_{\mathsf{MBI}}(y) := \frac{\exp(\log \nu_l(\eta,x,y) - \log \nu_l(\eta,x,x) + C_{\mathsf{MBI}})}{\exp(-\frac{d(x,y)^2}{2(s\eta)})} \le 1, \qquad \forall y \in B_x(r),$$

$$V_{\mathsf{RHK}}(x) := \frac{\exp(-f(x) + \log \nu_l(\eta,x,y) + f(y) - \log \nu_l(\eta,y,y) + C_{\mathsf{RHK}})}{\exp(-\frac{1}{2t}d(x,y)^2)} \le 1, \quad \forall x \in B_y(r).$$

Lemma 38 (Nowak et al. [2019], Nowak [2025])**.** *Let d denote the dimension, and ν denote the heat kernel (that corresponds to $\frac{\partial}{\partial t}\nu(t,x,y) = \frac{1}{2}\Delta\nu(t,x,y)$). For φ represent the size of neighborhood, we have*

$$\frac{c(\varphi,d)}{t^{\frac{d}{2}}}\exp(-\frac{1}{2t}d(x,y)^2) \le \nu(t,x,y) \le \frac{C(\varphi,d)}{t^{\frac{d}{2}}}\exp(-\frac{1}{2t}d(x,y)^2), \forall d(x,y) \le \varphi$$

Fix $t = \mathcal{O}(\frac{1}{d})$. When $\varphi = \mathcal{O}(\frac{1}{d})$, it holds that $\frac{C(\varphi,d)}{c(\varphi,d)} = \mathcal{O}(1)$. In this case, we denote $C = C(\varphi,d)$ and $c = c(\varphi,d)$.

Proposition 39. *Let $M = \mathcal{S}^d$ be a hypersphere. We set $\frac{1}{\eta} = L_1^2 d \log \frac{1}{\zeta}$ as the step size of proximal sampler, $\frac{1}{t} = L_1^2(d-2)\log\frac{1}{\zeta}$ as the parameter for proposal (Riemannian Gaussian) distribution of rejection sampling, and truncation level $l = \mathrm{Poly}\,(d^2\,\mathrm{Poly}\,\log\frac{1}{\zeta})$. There exists parameters*

$$C_{\mathsf{MBI}} = -\frac{d}{2}\log(2\pi) - \log(C+1), \quad C_{\mathsf{RHK}} = -\frac{1}{2\log\frac{1}{\zeta}} - \frac{d}{2}\log(2\pi) - \log(C+1)$$

s.t. $\overline{\pi}^{X|Y}, \overline{\pi}^{Y|X}$ satisfy Assumption 1.

Proof. See Proposition 42 and Proposition 43. □

Lemma 40. *[Xu, 2019, Equation 17] Let (M^d, g) be a complete manifold with Ricci curvature being non-negative. Then we have*

$$\frac{1}{(2\pi t)^{\frac{d}{2}}}\exp(-\frac{d(x,y)^2}{2t}) \le \nu(t,x,y)$$

I.1.1 Proof of Proposition 39

It's important to establish the order of a constant in algorithm first.

Lemma 41. *There exists some $C_{\mathsf{MBI}} = -\frac{d}{2}\log(2\pi) - \log(C+1)$ s.t.*

$$-\log\nu_l(\eta,x,y) + \log\nu_l(\eta,x,x) \ge \frac{d(x,y)^2}{2(s\eta)} + C_{\mathsf{MBI}}$$

Consequently, for $r = \sqrt{\frac{4-\frac{2}{s}}{C_\eta}}$, we have

$$V_{\mathsf{MBI}}(y) := \frac{\exp(\log \nu_l(\eta,x,y) - \log \nu_l(\eta,x,x) + C_{\mathsf{MBI}})}{\exp(-\frac{d(x,y)^2}{2(s\eta)})} \le 1, \forall y \in B_x(r).$$

Proof. Recall that we require $\frac{1}{\eta} = \tilde{\mathcal{O}}(\log\frac{1}{\zeta})$.

Write $\frac{1}{\eta} = C_\eta \log\frac{1}{\zeta}$ where $C_\eta = L_1^2 d$. Then we can write $e^{-\frac{1}{\eta C_\eta}} = \zeta$.

1. **Step 1** Consider neighborhood $B_x(r_0)$ with $r_0^2 = \frac{2}{C_\eta}$. We have

$$\nu_l(\eta, x, y) \le \nu(\eta, x, y) + \zeta \le \frac{C}{\eta^{\frac{d}{2}}} \exp(-\frac{d(x,y)^2}{2\eta}) + e^{-\frac{1}{\eta C_\eta}} = \frac{C}{\eta^{\frac{d}{2}}} \exp(-\frac{d(x,y)^2}{2\eta})(1 + \delta(x,\eta))$$

with $\delta(x,\eta) := \frac{e^{-\frac{1}{\eta C_\eta}}}{\frac{C}{\eta^{\frac{d}{2}}} \exp(-\frac{d(x,y)^2}{2\eta})}$. For this δ, we can see that

$$\delta(x,\eta) = \frac{e^{-\frac{1}{\eta C_\eta}}}{\frac{C}{\eta^{\frac{d}{2}}} \exp(-\frac{d(x,y)^2}{2\eta})} = \eta^{\frac{d}{2}} \frac{\exp(\frac{1}{\eta}(\frac{d(x,y)^2}{2} - \frac{1}{C_\eta}))}{C} \le \frac{1}{C}, \forall y \in B_x(r_0),$$

Hence $C(1 + \delta(x,\eta)) \le C + 1$. which further implies

$$\begin{aligned} -\log \nu_l(\eta, x, y) &\ge -\log C + \frac{d}{2}\log\eta + \frac{d(x,y)^2}{2\eta} - \log(1 + \delta(x,\eta)) \\ &\ge -\log(C+1) + \frac{d}{2}\log\eta + \frac{d(x,y)^2}{2s\eta}. \end{aligned}$$

Now consider a slightly larger neighborhood where $s > 1$ will be set later: $\frac{1}{C_\eta} \le \frac{d(x,y)^2}{2} \le \frac{2-\frac{1}{s}}{C_\eta}$, we have

$$\delta(x,\eta) = \frac{e^{-\frac{1}{\eta C_\eta}}}{\frac{C}{\eta^{\frac{d}{2}}} \exp(-\frac{d(x,y)^2}{2\eta})} \le \eta^{\frac{d}{2}} \frac{\exp(\frac{1-\frac{1}{s}}{\eta C_\eta})}{C} \le \frac{1}{C\zeta^{1-\frac{1}{s}}},$$

so that when ζ is small, we have

$$\begin{aligned} &\log(1 + \delta(x,\eta)) \approx \log\delta(x,\eta) \le \log\frac{1}{C} + (1 - \frac{1}{s})\log\frac{1}{\zeta} \\ =&\log\frac{1}{C} + \frac{1-\frac{1}{s}}{\eta C_\eta} \le \log\frac{1}{C} + \frac{1-\frac{1}{s}}{\eta}\frac{d(x,y)^2}{2}, \forall \frac{1}{C_\eta} \le \frac{d(x,y)^2}{2} \le \frac{2-\frac{1}{s}}{C_\eta}, \end{aligned}$$

which further implies

$$\begin{aligned} -\log \nu_l(\eta, x, y) &\ge -\log C + \frac{d}{2}\log\eta + \frac{d(x,y)^2}{2\eta} - \log(1 + \delta(x,\eta)) \\ &\ge -\log C + \frac{d}{2}\log\eta + \frac{d(x,y)^2}{2\eta} - \log\frac{1}{C} - \frac{1-\frac{1}{s}}{\eta}\frac{d(x,y)^2}{2} \\ &\ge \frac{d}{2}\log\eta + \frac{d(x,y)^2}{2(s\eta)} \\ &\ge -\log(C+1) + \frac{d}{2}\log\eta + \frac{d(x,y)^2}{2(s\eta)}. \end{aligned}$$

Together, we conclude that for all $y \in B_x(r)$,

$$-\log \nu_l(\eta, x, y) \ge -\log(C+1) + \frac{d}{2}\log\eta + \frac{d(x,y)^2}{2(s\eta)}.$$

2. **Step 2**
We have $\nu(\eta, x, x) \ge \frac{1}{(2\pi\eta)^{\frac{d}{2}}}$ and consequently

$$\nu_l(\eta, x, x) \ge \frac{1}{(2\pi\eta)^{\frac{d}{2}}} - \zeta \approx \frac{1}{(2\pi\eta)^{\frac{d}{2}}}.$$

Thus for all $y \in B_x(r)$, for some constant C we have

$$\begin{aligned} -\log \nu_l(\eta, x, y) + \log \nu_l(\eta, x, x) &\ge -\log(C+1) + \frac{d}{2}\log\eta + \frac{d(x,y)^2}{2(s\eta)} - \frac{d}{2}\log\eta - \frac{d}{2}\log(2\pi) \\ &= \frac{d(x,y)^2}{2(s\eta)} - \frac{d}{2}\log(2\pi) - \log(C+1) \end{aligned}$$

Therefore there exists some $C_{\mathsf{MBI}} = -\frac{d}{2}\log(2\pi) - \log(C+1)$ s.t.

$$V_{\mathsf{MBI}}(y) := \frac{\exp(\log \nu_l(\eta, x, y) - \log \nu_l(\eta, x, x) + C_{\mathsf{MBI}})}{\exp(-\frac{d(x,y)^2}{2(s\eta)})} \leq 1, \forall y \in B_x(r).$$

□

Proposition 42. *Let M be hypersphere $\mathcal{S}^d$ so that the truncation error bound can be proved in L_∞. Consider Algorithm 3 with $t = \eta s$ where $s = \frac{d}{d-2} > 1$ is a constant that does not depend on η, ζ. For small ζ, the error for inexact rejection sampling with ν_l is of order ζ, i.e., $\|\hat{\pi}^{Y|X} - \overline{\pi}^{Y|X}\|_{TV} = \tilde{\mathcal{O}}(\zeta)$. Hence by triangle inequality, $\|\pi^{Y|X} - \overline{\pi}^{Y|X}\|_{TV} = \tilde{\mathcal{O}}(\zeta)$.*

Proof. [Proof of Proposition 42]

Now we have for all $y \in B_x(r)$,

$$-\log \nu_l(\eta, x, y) + \log \nu_l(\eta, x, x) \geq \frac{d(x,y)^2}{2(s\eta)} + C_{\mathsf{MBI}}.$$

This implies,

$$\exp(\log \nu_l(\eta, x, y) - \log \nu_l(\eta, x, x) + C_{\mathsf{MBI}}) \leq \exp(-\frac{d(x,y)^2}{2(s\eta)}).$$

Hence we have

$$V_{\mathsf{MBI}}(y) := \frac{\exp(\log \nu_l(\eta, x, y) - \log \nu_l(\eta, x, x) + C_7)}{\exp(-\frac{d(x,y)^2}{2(s\eta)})} \leq 1, \forall y \in B_x(r).$$

Recall that μ denote the density for Riemannian Gaussian distribution. We compute

$$\begin{aligned}
\mu(s\eta, x, y)\frac{V_{\mathsf{MBI}}(y)}{\mathbb{E}_{\mu(s\eta,x,y)}V_{\mathsf{MBI}}(y)} =& \frac{V_{\mathsf{MBI}}(y)}{\int_M V_{\mathsf{MBI}}(y)\mu(s\eta, x, y)dV_g(y)}\mu(s\eta, x, y) \\
=& \frac{V_{\mathsf{MBI}}(y)\mu(s\eta, x, y)}{\int_M \frac{\exp(\log \nu_l(\eta,x,y) - \log \nu_l(\eta,x,x) + C_7)}{\exp(-\frac{d(x,y)^2}{2(s\eta)})}\mu(s\eta, x, y)dV_g(y)} \\
=& \frac{\exp(\log \nu_l(\eta, x, y) - \log \nu_l(\eta, x, x) + C_7)}{\int_M \exp(\log \nu_l(\eta, x, y) - \log \nu_l(\eta, x, x) + C_7)dV_g(y)} \\
=& \frac{\nu_l(\eta, x, y)}{\int_M \nu_l(\eta, x, y)dV_g(y)} =: \hat{\pi}^{Y|X}.
\end{aligned}$$

Thus the desired rejection sampling output can be written as

$$\hat{\pi}^{Y|X} = \mu(s\eta, x, y)\frac{V_{\mathsf{MBI}}(y)}{\mathbb{E}_{\mu(s\eta,x,y)}V_{\mathsf{MBI}}(y)}.$$

On the other hand we denote $\overline{V_{\mathsf{MBI}}}(y) = \min\{1, V_{\mathsf{MBI}}(y)\}$, and the actual rejection sampling output is $\overline{\pi}^{Y|X} = \mu(s\eta, x, y)\frac{\overline{V_{\mathsf{MBI}}}(y)}{\mathbb{E}_{\mu(s\eta,x,y)}\overline{V_{\mathsf{MBI}}}(y)}$. Following Fan et al. [2023, Proof of Theorem 6], we get

$$\begin{aligned}
\|\hat{\pi}^{Y|X} - \overline{\pi}^{Y|X}\|_{TV} &\leq \mathbb{E}_{\exp(-\frac{d(x,y)^2}{2(s\eta)})}[|\frac{V_{\mathsf{MBI}}}{\mathbb{E}[V_{\mathsf{MBI}}]} - \frac{\overline{V_{\mathsf{MBI}}}}{\mathbb{E}[\overline{V_{\mathsf{MBI}}}]}|] \\
&\leq \frac{2\mathbb{E}[|V_{\mathsf{MBI}} - \overline{V_{\mathsf{MBI}}}|]}{|\mathbb{E}[\overline{V_{\mathsf{MBI}}}]|}.
\end{aligned}$$

We aim to derive an upper bound for $\frac{2\mathbb{E}[|V_{\mathsf{MBI}} - \overline{V_{\mathsf{MBI}}}|]}{|\mathbb{E}[\overline{V_{\mathsf{MBI}}}]|}$. Note that

$$\mathbb{E}[|\overline{V_{\mathsf{MBI}}}|] \geq \int_{B_x(r)} \frac{\exp(\log \nu_l(\eta, x, y) - \log \nu_l(\eta, x, x) + C_{\mathsf{MBI}})}{\exp(-\frac{d(x,y)^2}{2(s\eta)})} \frac{\exp(-\frac{d(x,y)^2}{2(s\eta)})}{\int_M \exp(-\frac{d(x,y)^2}{2(s\eta)})dV_g(y)} dV_g(y)$$

$$= \frac{1}{\int_M \exp(-\frac{d(x,y)^2}{2(s\eta)})dV_g(y)} \int_{B_x(r)} \exp(\log \nu_l(\eta,x,y) - \log \nu_l(\eta,x,x) + C_{\mathsf{MBI}})dV_g(y).$$

and similarly

$$\mathbb{E}[|V_{\mathsf{MBI}} - \overline{V_{\mathsf{MBI}}}|] = \mathbb{E}_{\mu(s\eta,x,y)}[\frac{\exp(\log \nu_l(\eta,x,y) - \log \nu_l(\eta,x,x) + C_{\mathsf{MBI}})}{\exp(-\frac{d(x,y)^2}{2(s\eta)})} 1_{V_{\mathsf{MBI}}(y)>1}]$$

$$= \frac{1}{\int_M \exp(-\frac{d(x,y)^2}{2(s\eta)})dV_g(y)} \int_{\{V_{\mathsf{MBI}}(y)>1\}} \exp(\log \nu_l(\eta,x,y) - \log \nu_l(\eta,x,x) + C_{\mathsf{MBI}})dV_g(y).$$

Hence we only need to bound

$$\begin{aligned}
\frac{2\mathbb{E}[|V_{\mathsf{MBI}} - \overline{V_{\mathsf{MBI}}}|]}{|\mathbb{E}[\overline{V_{\mathsf{MBI}}}]|} &\leq 2\frac{\int_{\{V_{\mathsf{MBI}}(y)>1\}} \exp(\log \nu_l(\eta,x,y) - \log \nu_l(\eta,x,x) + C_{\mathsf{MBI}})dV_g(y)}{\int_{B_x(r)} \exp(\log \nu_l(\eta,x,y) - \log \nu_l(\eta,x,x) + C_{\mathsf{MBI}})dV_g(y)} \\
&= 2\frac{\int_{\{V_{\mathsf{MBI}}(y)>1\}} \frac{\nu_l(\eta,x,y)}{\nu_l(\eta,x,x)}dV_g(y)}{\int_{B_x(r)} \frac{\nu_l(\eta,x,y)}{\nu_l(\eta,x,x)}dV_g(y)} = 2\frac{\int_{\{V_{\mathsf{MBI}}(y)>1\}} \nu_l(\eta,x,y)dV_g(y)}{\int_{B_x(r)} \nu_l(\eta,x,y)dV_g(y)} \\
&\leq 2\frac{\int_{\{V_{\mathsf{MBI}}(y)>1\}} \zeta + \text{Poly}\,(\frac{1}{\eta})\exp(-\frac{d(x,y)^2}{2\eta})dV_g(y)}{\int_{B_x(r)} \nu_l(\eta,x,y)dV_g(y)} \\
&\leq 2\frac{\int_{\{\frac{d(x,y)^2}{2} > \frac{2-\frac{1}{s}}{C_\eta}\}} \zeta + \text{Poly}\,(\frac{1}{\eta})\exp(-\frac{d(x,y)^2}{2\eta})dV_g(y)}{\int_{B_x(r)} \nu_l(\eta,x,y)dV_g(y)} = \mathcal{O}(\zeta)
\end{aligned}$$

We used $\nu_l(\eta,x,y) \leq \zeta + \nu(\eta,x,y)$ In the last equality, note that when $\frac{d(x,y)^2}{2} > \frac{2-\frac{1}{s}}{C_\eta}$, it holds that

$$\exp(-\frac{d(x,y)^2}{2\eta}) \leq \exp(-\frac{2-\frac{1}{s}}{\eta C_\eta}) = \zeta^{2-\frac{1}{s}}$$

and then we know $\zeta^{2-\frac{1}{s}}\,\text{Poly}\,(\frac{1}{\eta}) = \mathcal{O}(\zeta)$ because for small ζ, the term $\zeta^{1-\frac{1}{s}}\,\text{Poly}\,(\frac{1}{\eta}) = \mathcal{O}(1)$.

We used lower bound $\int_{B_x(r)} \nu_l(\eta,x,y)dV_g(y)$:

$$\int_{B_x(r)} \nu(\eta,x,y)dV_g(y) \geq \int_{B_x(r)} \frac{1}{(2\pi\eta)^{\frac{d}{2}}} \exp(-\frac{d(x,y)^2}{2\eta})dV_g(y)$$

and for the choice of r, we know this is lower bounded by a constant.

□

Proposition 43. *Let M be hypersphere $\mathcal{S}^d$. Consider Algorithm 2 with $\frac{1}{\eta} = L_1^2 d \log\frac{1}{\zeta}$ and $\frac{1}{t} = L_1^2(d-2)\log\frac{1}{\zeta}$. For small ζ, the error for inexact rejection sampling with ν_l is of order ζ, i.e., $\|\hat{\pi}^{X|Y} - \overline{\pi}^{X|Y}\|_{TV} = \tilde{\mathcal{O}}(\zeta)$.*

Proof. [Proof of Proposition 43]

1. **Step 1** Set $s = \frac{d}{d-1}$ in Lemma 41. Let $\frac{1}{\eta} = L_1^2 d\log\frac{1}{\zeta}$, $C_\eta = L_1^2 d$. Note that we have $r^2/2 = \frac{2-\frac{d-1}{d}}{L_1^2 d} = \frac{d+1}{L_1^2 d^2}$ and $\frac{1}{t} = L_1^2(d-2)\log\frac{1}{\zeta}$. Note that $t = \frac{d}{d-2}\eta$. We know, for all $x \in B_r(y)$, we have

$$-\log \nu_l(\eta,x,y) + \log \nu_l(\eta,y,y) \geq \frac{d(x,y)^2}{2(s\eta)} + C_{\mathsf{MBI}}.$$

We want to find C_{RHK} so that the previously defined $t = \frac{d}{d-2}\eta$ can be the variable for proposal distribution, i.e., we need $f(x) - \log \nu_l(\eta,x,y) - f(y) + \log \nu_l(\eta,y,y) \geq \frac{1}{2t}d(x,y)^2 + C_{\mathsf{RHK}}$ to hold for all $x \in B_r(y)$. Hence we require

$$\frac{d(x,y)^2}{2(\frac{d}{d-1}\eta)} + C_{\mathsf{MBI}} - \frac{d(x,y)^2}{2t} - L_1 d(x,y) \geq C_{\mathsf{RHK}}$$

Note that

$$\begin{aligned}&\frac{d(x,y)^2}{2(\frac{d}{d-1}\eta)}-\frac{d(x,y)^2}{2t}-L_1d(x,y)+C_{\mathsf{MBI}}\\=&\frac{d(x,y)^2}{2}(L_1^2\log\frac{1}{\zeta})-L_1d(x,y)+C_{\mathsf{MBI}}\geq-\frac{1}{2\log\frac{1}{\zeta}}+C_{\mathsf{MBI}},\end{aligned}$$

where in the last inequality we take $d(x,y)=\frac{1}{L_1\log\frac{1}{\zeta}}$. Also note that when ζ is small, $|-\frac{1}{2\log\frac{1}{\zeta}}|$ is small. We can just take $C_{\mathsf{RHK}}=-\frac{1}{2\log\frac{1}{\zeta}}+C_{\mathsf{MBI}}\approx C_{\mathsf{MBI}}$.

Hence there exists constant $C_{\mathsf{RHK}}=-\frac{1}{2\log\frac{1}{\zeta}}-\frac{d}{2}\log(2\pi)-\log(C+1)$ s.t. for all $x\in B_y(r)$,

$$f(x)-\log\nu_l(\eta,x,y)-f(y)+\log\nu_l(\eta,y,y)-C_{\mathsf{RHK}}\geq\frac{1}{2t}d(x,y)^2.$$

2. **Step 2**

Denote

$$V_{\mathsf{RHK}}(x)=\frac{\exp(-f(x)+\log\nu_l(\eta,x,y)+f(y)-\log\nu_l(\eta,y,y)+C_{\mathsf{RHK}})}{\exp(-\frac{1}{2t}d(x,y)^2)}$$

and $\overline{V}_{\mathsf{RHK}}(x)=\min\{1,V_{\mathsf{RHK}}(x)\}$. Recall that the desired rejection sampling output can be written as $\hat{\pi}^{X|Y}=\mu(t,x,y)\frac{V_{\mathsf{RHK}}(x)}{\mathbb{E}_{\mu(t,x,y)}V_{\mathsf{RHK}}(x)}$. On the other hand the actual rejection sampling output is $\overline{\pi}^{X|Y}=\mu(t,x,y)\frac{\overline{V_{\mathsf{RHK}}}(x)}{\mathbb{E}_{\mu(t,x,y)}\overline{V_{\mathsf{RHK}}}(x)}$. Following Fan et al. [2023, Proof of Theorem 6], we get

$$\begin{aligned}\|\hat{\pi}^{X|Y}-\overline{\pi}^{X|Y}\|_{TV}&\leq\mathbb{E}_{\exp(-\frac{d(x,y)^2}{2t})}[|\frac{V_{\mathsf{RHK}}}{\mathbb{E}[V_{\mathsf{RHK}}]}-\frac{\overline{V_{\mathsf{RHK}}}}{\mathbb{E}[\overline{V_{\mathsf{RHK}}}]}|]\\&\leq\frac{2\mathbb{E}[|V_{\mathsf{RHK}}-\overline{V_{\mathsf{RHK}}}|]}{|\mathbb{E}[\overline{V_{\mathsf{RHK}}}]|}.\end{aligned}$$

we only need to bound

$$\begin{aligned}\frac{2\mathbb{E}[|V_{\mathsf{RHK}}-\overline{V_{\mathsf{RHK}}}|]}{|\mathbb{E}[\overline{V_{\mathsf{RHK}}}]|}&\leq2\frac{\int_{\{V_{\mathsf{RHK}}(y)>1\}}\exp(-f(x)+f(y)+\log\nu_l(\eta,x,y)-\log\nu_l(\eta,y,y)+C_{\mathsf{RHK}})dV_g(x)}{\int_{B_x(r)}\exp(-f(x)+f(y)+\log\nu_l(\eta,x,y)-\log\nu_l(\eta,y,y)+C_{\mathsf{RHK}})dV_g(x)}\\&=2\frac{\int_{\{V_{\mathsf{RHK}}(y)>1\}}\exp(-f(x)+f(y)+\log\nu_l(\eta,x,y))dV_g(x)}{\int_{B_x(r)}\exp(-f(x)+f(y)+\log\nu_l(\eta,x,y))dV_g(x)}\\&\leq2\frac{\int_{\{V_{\mathsf{RHK}}(y)>1\}}\exp(L_1d(x,y))\Big(\nu(\eta,x,y)+\zeta\Big)dV_g(x)}{\int_{B_x(r)}\exp(-L_1d(x,y))\nu_l(\eta,x,y)dV_g(x)}\end{aligned}$$

So it suffices to upper bound

$$\int_{\{V_{\mathsf{RHK}}(y)>1\}}\exp(L_1d(x,y))\nu(\eta,x,y)dV_g(x)$$

We need a sharper bound for distant points. With $\frac{1}{T}=L_1^2(d-0.5)\log\frac{1}{\zeta}$, we have

$$\begin{aligned}&-f(x)+f(y)+\log\nu(\eta,x,y)+\frac{1}{2T}d(x,y)^2\\\leq&L_1d(x,y)+\log\mathrm{Poly}\,(\frac{1}{\eta})-\frac{d(x,y)^2}{2\eta}+\frac{1}{2}d(x,y)^2(L_1^2(d-0.5)\log\frac{1}{\zeta})\\\leq&L_1d(x,y)+\log\mathrm{Poly}\,(\frac{1}{\eta})-\frac{d(x,y)^2}{2}\frac{1}{2}L_1^2\log\frac{1}{\zeta}\\\leq&\frac{1}{\log\frac{1}{\zeta}}+\log\mathrm{Poly}\,(\frac{1}{\eta}),\end{aligned}$$

where in the last inequality we set $d(x,y) = \frac{2}{L_1 \log \frac{1}{\zeta}}$

Hence

$$\begin{aligned}
&\int_{\{V_{\mathsf{RHK}}(y)>1\}} \exp(Ld(x,y))\nu(\eta,x,y)dV_g(x) \\
\leq &\int_{\{V_{\mathsf{RHK}}(y)>1\}} \exp(\frac{1}{\log\frac{1}{\zeta}} + \log \mathrm{Poly}\,(\frac{1}{\eta}) - \frac{1}{2T}d(x,y)^2)dV_g(x) \\
\leq &\,\mathrm{Poly}\,(\frac{1}{\eta}) \int_{\{V_{\mathsf{RHK}}(y)>1\}} \exp(-\frac{1}{2T}d(x,y)^2)dV_g(x) \\
\leq &\,\mathrm{Poly}\,(\frac{1}{\eta}) \exp(-(\frac{d+1}{d})(\frac{d-0.5}{d}\log\frac{1}{\zeta})) \\
= &\mathcal{O}(\mathrm{Poly}\,(\frac{1}{\eta})\zeta^{\frac{d^2+0.5d-0.5}{d^2}}) = \mathcal{O}(\zeta)
\end{aligned}$$

Here we used the fact that $\frac{d^2+0.5d-0.5}{d^2} > 1$.

□

Proposition 44. *Let $R = \frac{\sqrt{2}}{L\sqrt{d}}$. The cost for rejection sampling as in Proposition 39 is $\mathcal{O}(1)$ number of rejections, in ζ and dimension.*

Proof. [Proof of Proposition 44]

Step 1: We first show that in a local neighborhood we have

$$-\log \nu_l(\eta,x,y) + \log \nu_l(\eta,x,x) \leq \frac{d(x,y)^2}{2\eta} + Const$$

We have that

$$\begin{aligned}
\nu_l(\eta,x,y) \geq \nu(\eta,x,y) - \zeta &\geq \frac{c}{\eta^{\frac{d}{2}}}\exp(-\frac{d(x,y)^2}{2\eta}) - e^{-\frac{1}{\eta C_\eta}} \\
&= \frac{c}{\eta^{\frac{d}{2}}}\exp(-\frac{d(x,y)^2}{2\eta})(1-\delta(x,\eta))
\end{aligned}$$

where $C_\eta = L_1^2 d$ so that $\frac{1}{\eta} = C_\eta \log\frac{1}{\zeta}$.

We have that for $d(x,y)^2 \leq \frac{2}{C_\eta}$,

$$\delta(x,\eta) = \frac{e^{-\frac{1}{\eta C_\eta}}}{\frac{c}{\eta^{\frac{d}{2}}}\exp(-\frac{d(x,y)^2}{2\eta})} = \eta^{\frac{d}{2}}\frac{\exp(\frac{1}{\eta}(\frac{d(x,y)^2}{2} - \frac{1}{C_\eta}))}{c} \leq \frac{\eta^{\frac{d}{2}}}{c}.$$

Thus we can assume W.L.O.G. that $\delta(x,\eta) < C_\delta < 1$ is small enough s.t. $\log(1-\delta(x,\eta))$ is of constant order, for all $d(x,y)^2 \leq \frac{2}{C_\eta}$. Then we have

$$\begin{aligned}
-\log \nu_l(\eta,x,y) &\leq -\log c + \frac{d}{2}\log\eta + \frac{d(x,y)^2}{2\eta} - \log(1-\delta(x,\eta)) \\
&\leq \frac{d}{2}\log\eta + \frac{d(x,y)^2}{2\eta} - \log c, \quad \forall d(x,y)^2 \leq \frac{2}{C_\eta}
\end{aligned}$$

On the other hand, when $x = y$, we have $\nu(\eta,x,x) \leq \frac{C}{\eta^{\frac{d}{2}}}$.

$$\nu_l(\eta,x,x) \leq \frac{C}{\eta^{\frac{d}{2}}} + \zeta = \frac{C}{\eta^{\frac{d}{2}}} + e^{-\frac{1}{\eta C_\eta}} \lesssim \frac{C}{\eta^{\frac{d}{2}}},$$

observing that $e^{-\frac{1}{\eta C_\eta}} = \zeta \le C_\eta \log \frac{1}{\zeta} = \frac{1}{\eta}$. Thus we have

$$\log \nu_l(\eta, x, x) \le \log C - \frac{d}{2} \log \eta.$$

Hence for C, c we have

$$\begin{aligned}
&- \log \nu_l(\eta, x, y) + \log \nu_l(\eta, x, x) \\
\le &\frac{d}{2} \log \eta + \frac{d(x,y)^2}{2\eta} - \log c + \log C - \frac{d}{2} \log \eta = \frac{d(x,y)^2}{2\eta} + \log \frac{C}{c}, \quad \forall d(x,y)^2 \le \frac{2}{C_\eta}.
\end{aligned}$$

Step 2: This step is similar to Proposition 48. With the first step, we see that

$$f(x) - f(y) - \log \nu_l(\eta, x, y) + \log \nu_l(\eta, x, x) - C_{\mathsf{RHK}} \le L_1 d(x,y) + \frac{d(x,y)^2}{2\eta} + \log \frac{C}{c}$$

where $C_{\mathsf{RHK}} = -\frac{d}{2} \log(2\pi) - \log(C+1) - \frac{1}{2 \log \frac{1}{\zeta}}$, and c can be taken as $\frac{1}{(2\pi)^{\frac{d}{2}}}$.

With $\eta = \frac{1}{L_1^2 d \log \frac{1}{\zeta}}$ and $T = \frac{1}{L_1^2 (d+1) \log \frac{1}{\zeta}}$, viewing left hand side as a quadratic function of $d(x,y)$ we get

$$L_1 d(x,y) + \frac{d(x,y)^2}{2\eta} - \frac{d(x,y)^2}{2T} \le \frac{1}{2 \log \frac{1}{\zeta}}$$

Thus we have

$$\begin{aligned}
&f(x) - f(y) - \log \nu_l(\eta, x, y) + \log \nu_l(\eta, x, x) - C_{\mathsf{RHK}} \\
\le &L_1 d(x,y) + \frac{d(x,y)^2}{2\eta} + \log \frac{C}{c} \le \frac{1}{2 \log \frac{1}{\zeta}} + \frac{d(x,y)^2}{2T} + \log \frac{C}{c}
\end{aligned}$$

The neighborhood for which the bound is valid is $d(x,y)^2 \le \frac{2}{C_\eta}$, i.e., a ball with radius $R = \frac{\sqrt{2}}{L\sqrt{d}}$. We have that $(\frac{R}{\sin R})^{d-1} = \mathcal{O}(1)$. This allows us to do as exactly in Proposition 49 (with t from Proposition 39) to show the rejection sampling procedure finishes with $\mathcal{O}(1)$ number of rejections. □

Proof. [Proof of Corollary 10] Using Pinsker's inequality, we have

$$\|\rho_k^X - \pi^X\|_{TV} \le \sqrt{\frac{1}{2} H_{\pi^X}(\rho_k^X)} \le \sqrt{\frac{1}{2} \frac{H_{\pi^X}(\rho_0^X)}{(1+\eta\alpha)^{2k}}} \le \frac{1}{2}\varepsilon.$$

We want to bound $\|\rho_k^X - \pi^X\|_{TV} \le \frac{1}{2}\varepsilon$. It suffices to have $\frac{H_{\pi^X}(\rho_0^X)}{(1+\eta\alpha)^{2k}} \le \frac{1}{2}\varepsilon^2$. Hence we need $\log(\frac{2H_{\pi^X}(\rho_0^X)}{\varepsilon^2}) \le 2k \log(1+\eta\alpha)$, i.e., $k = \mathcal{O}\left(\frac{\log \frac{H_{\pi^X}(\rho_0^X)}{\varepsilon^2}}{\log(1+\eta\alpha)}\right)$.

For small step size η, we have $\frac{1}{\log(1+\eta\alpha)} = \mathcal{O}(\frac{1}{\eta\alpha})$. Hence $k = \mathcal{O}\left(\frac{1}{\eta\alpha} \log \frac{H_{\pi^X}(\rho_0^X)}{\varepsilon^2}\right) = \tilde{\mathcal{O}}(\frac{1}{\alpha\eta} \log \frac{1}{\varepsilon})$.

Using Proposition 44, by setting $\frac{1}{\eta} = L_1^2 d \log \frac{1}{\zeta}$, the expected number of rejections in rejection sampling is $\mathcal{O}(1)$. We pick $\zeta = \frac{\alpha\varepsilon}{L_1^2 d \log^2 \frac{1}{\varepsilon}}$ and consequently $\frac{1}{\eta} = L_1^2 d \log \frac{L_1^2 d \log^2 \frac{1}{\varepsilon}}{\alpha\varepsilon} = \tilde{\mathcal{O}}(L_1^2 d \log \frac{1}{\varepsilon})$. It follows that

$$k = \mathcal{O}(\frac{1}{\alpha\eta} \log \frac{1}{\varepsilon}) = \tilde{\mathcal{O}}(\frac{L_1^2 d}{\alpha} \log^2 \frac{1}{\varepsilon}).$$

The result then follows from triangle inequality:

$$\|\tilde{\rho}_k^X - \pi^X\|_{TV} \le \|\tilde{\rho}_k^X - \rho_k^X\|_{TV} + \|\rho_k^X - \pi^X\|_{TV} \le k(\zeta_{\mathsf{RHK}} + \zeta_{\mathsf{MBI}}) + \frac{1}{2}\varepsilon = \tilde{\mathcal{O}}(\varepsilon)$$

where from Proposition 39, we can set the ε to be $\frac{\zeta}{k}$, so that $k\zeta = \tilde{\mathcal{O}}\left(\frac{L^2 d}{\alpha}(\log^2 \frac{1}{\varepsilon}) \frac{\alpha\varepsilon}{L^2 d \log^2 \frac{1}{\varepsilon}}\right) = \tilde{\mathcal{O}}(\varepsilon)$. □

I.1.2 Heat kernel truncation: hypersphere

In this subsection, we show that on hyperspheres $\mathcal{S}^d$, the truncation error bound $\|\nu - \nu_L\|_\infty = \tilde{\mathcal{O}}(\zeta)$ can be achieved with truncation level $L = \tilde{\mathcal{O}}(\mathrm{Poly}\,(\log \frac{1}{\varepsilon}))$. As proved in Zhao and Song [2018], the heat kernel on $\mathcal{S}^d$ can be written as the following uniformly convergent series (with $\varphi := \langle x, y\rangle_{\mathbb{R}^{d+1}}$)

$$\nu(\eta, x, y) = \sum_{k=0}^{\infty} \exp(-\frac{k(k+d-1)t}{2}) \frac{2k+d-1}{(d-1)A_{\mathcal{S}^d}} C_k^{(d-1)/2}(\cos(\varphi)),$$

where C_l^α are the Gegenbauer polynomials. Define

$$M_l = \frac{\Gamma(\frac{l+d-1}{2})}{\Gamma(\frac{d-1}{2})\Gamma(\frac{l}{2}+1)} + \left| \frac{\Gamma(l+d-1)}{\Gamma(d-1)\Gamma(l+1)} - \frac{\Gamma(\frac{l+d-1}{2})}{\Gamma(\frac{d-1}{2})\Gamma(\frac{l}{2}+1)} \right|.$$

Such M_l is constructed to be an upper bound for Gegenbauer polynomials; see Zhao and Song [2018, Proof of Theorem 1]. The following proposition is directly implied by Zhao and Song [2018, Theorem 1], and we provide a proof for completeness.

Lemma 45. *For $l = \Theta(d^2)$, we have*

$$\frac{M_{l+1}}{M_l} \leq \frac{(l+d-1)^{d-2}}{(l+1)^{d-2}}$$

and consequently $\frac{M_{l+1}}{M_l} = \mathcal{O}(1)$.

Proof. For $l = \Theta(d^2)$, we have (by definition of M and Gamma function) $M_l = \frac{(l+d-2)!}{(d-2)!l!}$ and $M_{l+1} = \frac{(l+1+d-2)!}{(d-2)!(l+1)!}$. Hence $M_l = \frac{(l+d-2)!}{(d-2)!l!} \geq \frac{(l+1)^{d-2}}{(d-2)!}$ and $M_{l+1} = \frac{(l+1+d-2)!}{(d-2)!(l+1)!} \leq \frac{(l+d-1)^{d-2}}{(d-2)!}$ Then we have (note that $l = \Theta(d^2)$)

$$\frac{M_{l+1}}{M_l} \leq \frac{(l+d-1)^{d-2}}{(l+1)^{d-2}} = (1 + \frac{d-2}{l+1})^{d-2} = \mathcal{O}((1+\frac{1}{d})^d) = \mathcal{O}(1)$$

□

Proposition 46. *Let $M = \mathcal{S}^d$ be a hypersphere. For truncation level $L = \Theta(d^2\,\mathrm{Poly}\,(\log \frac{1}{\varepsilon}))$, we can achieve $|\nu(\eta, x, y) - \nu_L(\eta, x, y)| = \tilde{\mathcal{O}}(\zeta), \forall x, y \in \mathcal{S}^d$.*

Proof. [Proof of Proposition 46] Throughout the proof, we denote $\varphi = \langle x, y\rangle_{\mathbb{R}^{d+1}}$. The parameters M_l satisfies $|C_l^{\frac{d-2}{2}}(x)| \leq M_l$ according to Zhao and Song [2018, Proof of Theorem 1]. Hence, we have

$$\begin{aligned}
&|\nu(\eta, x, y) - \nu_L(\eta, x, y)| \\
=&|\sum_{l=L+1}^{\infty} \exp(-\frac{l(l+d-1)\eta}{2}) \frac{2l+d-1}{(d-1)A_{\mathcal{S}^d}} C_l^{(d-1)/2}(\cos(\varphi))| \\
\leq& \sum_{l=L+1}^{\infty} \exp(-\frac{l(l+d-1)\eta}{2}) \frac{(2l+d-1)M_l}{(d-1)A_{\mathcal{S}^d}}.
\end{aligned}$$

Observe that for all $l \geq L+1$, since for large L (that depends on dimension) we have $\frac{M_{l+1}}{M_l} = \mathcal{O}(1)$; see, also, Zhao and Song [2018, Proof of Theorem 1]. Hence,

$$\begin{aligned}
Q_l :=& \frac{\exp(-\frac{(l+1)(l+1+d-1)\eta}{2})(2l+d+1)M_{l+1}}{\exp(-\frac{l(l+d-1)\eta}{2})(2l+d-1)M_l} = \frac{2l+d+1}{2l+d-1} \exp(-\frac{(d+2l)\eta}{2}) \frac{M_{l+1}}{M_l} \\
=& \mathcal{O}(\exp(-\frac{(2l+d)\eta}{2}) \frac{M_{l+1}}{M_l}) \\
=& \mathcal{O}(\exp(-\frac{(2L+2+d)\eta}{2})) = \mathcal{O}(\exp(-L\eta)) = \mathcal{O}(\zeta).
\end{aligned}$$

Algorithm 6 Riemannian Gaussian on hypersphere through rejection sampling

Input $xx^* \in \mathcal{S}^d$. Require $t \le \frac{6}{(d+1)^2}$.
for $i = 0, 1, 2, ...$ **do**
  Generate proposal $v \propto e^{-\frac{1}{2t}\|v\|^2}$ (in Euclidean space), repeat until $\|v\| \le \pi$.
  Generate u uniformly on $[0, 1]$.
  Return v if $u \le \frac{e^{-\frac{1}{2t}|v|^2}(\frac{\sin|v|}{|v|})^{d-1}}{e^{-\frac{1}{2t}|v|^2}}$
end for
Generate E to be an orthonormal basis for $T_x\mathcal{S}^d$, set $v \leftarrow v \circ E \in T_x\mathcal{S}^d$.
Output sample $y = \exp_{x^*}(v)$

For the last line, note that with $L = \text{Poly}\,(\log\frac{1}{\zeta})$ and $\eta = \frac{1}{C\log\frac{1}{\varepsilon}}$, we have that $L\eta = \text{Poly}\,(\log\frac{1}{\zeta})$. This implies $\exp(-L\eta) = \mathcal{O}(\exp(\log\zeta)) = \mathcal{O}(\zeta)$. Now we compute the truncation error.

$$
\begin{aligned}
&|\nu(\eta, x, y) - \nu_L(\eta, x, y)| \\
=&|\sum_{l=L+1}^{\infty} \exp(-\frac{l(l+d-1)\eta}{2})\frac{2l+d-1}{(d-1)A_{\mathcal{S}^d}}C_l^{(d-1)/2}(\cos(\varphi))| \\
\le&\frac{1}{(d-1)A_{\mathcal{S}^d}}\sum_{l=L+1}^{\infty} \exp(-\frac{l(l+d-1)\eta}{2})(2l+d-1)M_l \\
\le&\frac{1}{(d-1)A_{\mathcal{S}^d}}\exp(-\frac{(L+1)(L+1+d-1)\eta}{2})(2L+d+1)\frac{(L+d-1)^{d-2}}{(d-2)!}\frac{1}{1-Q} \\
=&\tilde{\mathcal{O}}(\exp(-\frac{(L+1)(L+d)\eta}{2})(2L+d+1)\frac{(L+d-1)^{d-2}}{(d-2)!}) \\
=&\tilde{\mathcal{O}}(\exp(-L^2\eta)\,\text{Poly}\,(L)) = \tilde{\mathcal{O}}(\exp(-\,\text{Poly}\,(\log\frac{1}{\zeta}))\,\text{Poly}\,(\log\frac{1}{\zeta})) = \tilde{\mathcal{O}}(\zeta).
\end{aligned}
$$

□

I.2 Sampling from the Riemannian Gaussian distribution on hypersphere

We show that Riemannian Gaussian distribution on hypersphere can be generated efficiently through rejection sampling (Algorithm 6).

Note that Algorithm 6 first generate a Euclidean Gaussian v in the tangent space, which is guaranteed to satisfy $\|v\| \le \pi$. Its density can be computed exactly under the normal coordinates. Then we perform rejection sampling to get samples v s.t. $\exp_{x^*}(v)$ follow the Riemannian Gaussian distribution $\mu(t, x^*, \cdot)$ exactly.

Proposition 47. *On hypersphere $\mathcal{S}^d$, when $t = \mathcal{O}(\frac{1}{d^2})$, Algorithm 6 output a sample following Riemannian Gaussian distribution $\mu(t, x^*, y) \propto \exp(-\frac{1}{2t}d(x^*, y)^2)$, with iteration complexity $\mathcal{O}(1)$.*

Proof. First note that when we generate a tangent space Gaussian restricted to $\|x\| \le \pi$, under normal coordinates the corresponding density would be $\propto e^{-\frac{1}{2t}|x|^2}, x \in B_0(\pi)$.

Recall that the Riemannian metric g of $\mathcal{S}^d$ under normal coordinates satisfies $\sqrt{\det g} = (\frac{\sin|y|}{|y|})^{d-1}$. For Riemannian Gaussian distribution, we change it from Riemannian volume measure to the measure in local coordinates, and under the local coordinates, it has density

$$
\propto \exp(-\frac{1}{2t}|x|^2)\sqrt{\det g} = \exp(-\frac{1}{2t}|x|^2)(\frac{\sin|x|}{|x|})^{d-1}
$$

Therefore we can compute the number of expected rejections as

$$
\frac{\int_{B_0(\pi)} e^{-\frac{1}{2t}|x|^2}dx}{\int_{B_0(\pi)} e^{-\frac{1}{2t}|x|^2}(\frac{\sin|x|}{|x|})^{d-1}dx}
$$

Recall that with $R = \sqrt{\frac{6}{1+d}}$, we know $(\frac{R}{\sin R})^{d-1} = \mathcal{O}(1)$. Then we have

$$\int_{B_0(\pi)} e^{-\frac{1}{2t}|x|^2} (\frac{\sin|x|}{|x|})^{d-1} dx \geq (\frac{\sin R}{R})^{d-1} \int_{B_0(R)} e^{-\frac{1}{2t}|x|^2} dx$$

Then we get

$$\frac{\int_{B_0(\pi)} e^{-\frac{1}{2t}|x|^2} dx}{\int_{B_0(\pi)} e^{-\frac{1}{2t}|x|^2} (\frac{\sin|x|}{|x|})^{d-1} dx} \leq (\frac{R}{\sin R})^{d-1} \frac{\int_{B_0(\pi)} e^{-\frac{1}{2t}|x|^2} dx}{\int_{B_0(R)} e^{-\frac{1}{2t}|x|^2} dx}$$

$$= \mathcal{O}(\frac{(2\pi t)^{\frac{d}{2}}}{(2\pi t)^{\frac{d}{2}}(1 - \exp(-\frac{1}{2}(\frac{R^2}{t} - d)))}) = \mathcal{O}(1)$$

As long as $\frac{R^2}{t} - d \geq 1$, i.e, $t \leq \frac{6}{(1+d)^2}$, the last equality holds. □

I.3 Varadhan's asymptotics

We consider the approximation scheme introduced in Section A.2 using Varadhan's asymptotics. Let $\varphi(x) = \frac{1}{2\eta} d(x,y)^2$. Intuitively, we want to see how the function φ can improve the convexity of $f + \varphi$.

On a manifold with positive curvature, we consider the situation that we cannot compute the minimizer of $g(x) = f(x) + \frac{1}{2\eta} d(x,y)^2$, and instead use y as the approximation of it. Notice that when η is small, since $f(x)$ is uniformly bounded, the function $g(x)$ is dominated by $\frac{1}{2\eta} d(x,y)^2$, thus the minimizer of g will be close to y. Therefore it is reasonable to use y as an approximation of the mode of $e^{-g(x)}$. Then in rejection sampling, we use $\mu(t, y, x)$ as the proposal.

Let L_1 be the Lipschitz constant of f. In the next proposition, we show that for some constant C_ε, with certain choices of η and t, it holds that

$$f(x) + \frac{1}{2\eta} d(x,y)^2 - f(y) + \frac{C_\varepsilon}{2} \geq \frac{1}{2t} d(x,y)^2.$$

Consequently, the acceptance rate defined by

$$V(x) := \frac{\exp(-f(x) + f(y) - \frac{1}{2\eta} d(x,y)^2 - \frac{C_\varepsilon}{2})}{\exp(-\frac{1}{2t} d(x,y)^2)},$$

is guaranteed to be bounded by 1. Then, in Proposition 49 we show that the expected number of rejections is $\mathcal{O}(1)$ in dimension d and step size η.

Proposition 48. *Let f be L_1-Lipschitz and C_ε be some constant. Take $\eta = \frac{C_\varepsilon}{L_1^2 d}$. With $T = \frac{C_\varepsilon}{L_1^2(d+1)}$ and $t = \frac{C_\varepsilon}{L_1^2(d-1)}$, it holds that*

$$\frac{1}{2T} d(x,y)^2 + C_\varepsilon \geq f(x) + \frac{1}{2\eta} d(x,y)^2 - f(y) + \frac{C_\varepsilon}{2} \geq \frac{1}{2t} d(x,y)^2.$$

Consequently, the acceptance rate is bounded by 1, i.e., $V(x) \leq 1, \forall x \in M$.

Proof. [Proof of Proposition 48] Since f is L_1-Lipschitz, we have $\|\operatorname{grad} f(x)\| \leq L_1$. Then we have $L_1 d(x,y) \geq f(x) - f(y) \geq -L_1 d(x,y)$.

1. **The lower bound:** The goal is to find some $t > 0$ and constant C such that

$$f(x) + \frac{1}{2\eta} d(x,y)^2 - f(y) + C \geq \frac{1}{2t} d(x,y)^2.$$

It suffices to find t, C such that

$$\frac{1}{2\eta} d(x,y)^2 - \frac{1}{2t} d(x,y)^2 - L_1 d(x,y) + C \geq 0.$$

The left hand side can be viewed as a quadratic function of $d(x,y)$. When $d(x,y) = \frac{L_1}{\frac{1}{\eta}-\frac{1}{t}}$, the left hand side is minimized, and the mimimum is $-\frac{1}{2}\frac{L_1^2}{\frac{1}{\eta}-\frac{1}{t}} + C$. Hence we can take $C = \frac{1}{2}\frac{L_1^2}{\frac{1}{\eta}-\frac{1}{t}}$. Take $\eta = \frac{C_\varepsilon}{L_1^2 d}$ and $t = \frac{C_\varepsilon}{L_1^2(d-1)}$. Then we have $C = \frac{1}{2}\frac{L_1^2}{\frac{1}{\eta}-\frac{1}{t}} = \frac{C_\varepsilon}{2}$.

2. **The upper bound:** For an upper bound, we want some $T \leq \eta$ for which we want to show that

$$f(x) + \frac{1}{2\eta}d(x,y)^2 - f(y) - \frac{C_\varepsilon}{2} \leq L_1 d(x,y) + \frac{1}{2\eta}d(x,y)^2 - \frac{C_\varepsilon}{2} \leq \frac{1}{2T}d(x,y)^2.$$

Similar as before, it suffices to show

$$(\frac{1}{2\eta} - \frac{1}{2T})d(x,y)^2 + L_1 d(x,y) - \frac{C_\varepsilon}{2} \leq 0.$$

The left hand side is maximized at $d(x,y) = \frac{L_1}{\frac{1}{T}-\frac{1}{\eta}}$, with maximum $\frac{1}{2}\frac{L_1^2}{\frac{1}{T}-\frac{1}{\eta}} - \frac{C_\varepsilon}{2}$. Take $T = \frac{C_\varepsilon}{L_1^2(d+1)}$. We can then verify that

$$\frac{1}{2}\frac{L_1^2}{\frac{1}{T}-\frac{1}{\eta}} - \frac{C_\varepsilon}{2} = \frac{1}{2}\frac{L_1^2}{L_1^2/C_\varepsilon} - \frac{C_\varepsilon}{2} = 0.$$

3. **Combining the two steps:** From the above two steps, we get

$$\frac{1}{2T}d(x,y)^2 + C_\varepsilon \geq f(x) + \frac{1}{2\eta}d(x,y)^2 - f(y) + \frac{C_\varepsilon}{2} \geq \frac{1}{2t}d(x,y)^2.$$

□

In the following proposition, we show that on a hypersphere (where the Riemannian metric in normal coordinates is well studied), the expected number of rejections which equals to

$$\frac{\int_M \exp(-\frac{1}{2t}d(x,y)^2)dV_g(x)}{\int_M \exp(-f(x)+f(y)-\frac{1}{2\eta}d(x,y)^2 - \frac{C_\varepsilon}{2})dV_g(x)},$$

which is independent of dimension and accuracy.

Proposition 49. *Let M be hypersphere. Set $C_\varepsilon = \frac{1}{\log\frac{1}{\varepsilon}}$. Assume without loss of generality that $L_1 \geq \max\{1, \frac{d+1}{\sqrt{6}}\}$. Then with $\eta = \frac{C_\varepsilon}{L_1^2 d}$ and $t = \frac{C_\varepsilon}{L_1^2(d-1)}$, for small ε, the expected number of rejections is $\mathcal{O}(1)$ in both dimension and ε.*

Proof. Let $T = \frac{C_\varepsilon}{L_1^2(d+1)}$. We try to bound the expected number of rejections. We compute it as follows:

$$\frac{\int_M \exp(-\frac{1}{2t}d(x,y)^2)dV_g(x)}{\int_M \exp(-f(x)+f(y)-\frac{1}{2\eta}d(x,y)^2 - \frac{C_\varepsilon}{2})dV_g(x)} \leq \frac{\int_M \exp(-\frac{1}{2t}d(x,y)^2)dV_g(x)}{\int_M \exp(-\frac{1}{2T}d(x,y)^2 - C_\varepsilon)dV_g(x)}.$$

Using Li and Erdogdu [2023, Lemma 8.2] and Li and Erdogdu [2023, Lemma C.5], when $\beta \geq \frac{d}{R^2}$, using Riemannian normal coordinates we have the following lower bound on the integral:

$$\begin{aligned}\int_M \exp(-\frac{\beta}{2}d(x,y)^2)dV_g(x) &\geq \int_{B_y(R)} \exp(-\frac{\beta}{2}d(x,y)^2)dV_g(x) \\ &\geq (\frac{\sin R}{R})^{d-1}(\frac{2\pi}{\beta})^{\frac{d}{2}}(1-\exp(-\frac{1}{2}(\beta R^2 - d))),\end{aligned}$$

where $B_y(R)$ denote the geodesic ball centered at y with radius R.

On the other hand, we have

$$\int_M \exp(-\frac{t}{2}d(x,y)^2)dV_g(x) \leq \int_{B_\pi(0)} \exp(-\frac{t}{2}|x|^2)dx \leq (\frac{2\pi}{t})^{\frac{d}{2}}.$$

We next find a suitably small R which only depends on dimension, for which we have $\frac{R}{\sin R} \leq 1 + \frac{1}{d}$. Using Taylor series for $\sin(R)$, we have $\frac{R}{\sin R} \approx \frac{R}{R - \frac{R^3}{6}}$. Hence for $R^2 \leq \frac{6}{1+d}$, we have (approximately) $\frac{R}{\sin R} \leq 1 + \frac{1}{d}$. Consequently we set $R = \sqrt{\frac{6}{1+d}}$, and we know $(\frac{R}{\sin R})^{d-1} = \mathcal{O}(1)$.

Combining the bounds discussed previously, we have

$$\begin{aligned}
&\frac{\int_M \exp(-\frac{1}{2t} d(x,y)^2) dV_g(x)}{\int_M \exp(-\frac{1}{2T} d(x,y)^2 - C_\varepsilon) dV_g(x)} \\
\leq & e^{C_\varepsilon} (\frac{R}{\sin R})^{d-1} \frac{(2\pi t)^{\frac{d}{2}}}{(2\pi T)^{\frac{d}{2}} (1 - \exp(-\frac{1}{2}(\frac{L_1^2(d+1)}{C_\varepsilon} R^2 - d)))} \\
\leq & e^{C_\varepsilon + 1} (\frac{t}{T})^{\frac{d}{2}} \frac{1}{1 - \exp(-\frac{1}{2}(\frac{L_1^2(d+1)}{C_\varepsilon} R^2 - d))} = e^{C_\varepsilon + 1} (\frac{d+1}{d-1})^{\frac{d}{2}} \frac{1}{1 - \exp(-\frac{1}{2}(\frac{L_1^2(d+1)}{C_\varepsilon} R^2 - d))}.
\end{aligned}$$

For small ε, we have $C_\varepsilon \leq 1$. Since we assumed $L_1 \geq 1$ and $L_1^2 \geq \frac{d+1}{6}$, we have $1 - \exp(-\frac{1}{2}(\frac{L_1^2(d+1)}{C_\varepsilon} R^2 - d)) \geq 1 - \exp(-\frac{1}{2}(\frac{(d+1)^2}{6} \frac{6}{d+1} - d)) \geq 1 - \exp(-\frac{1}{2})$. As a result, we see that the expect number of rejections is of order $\mathcal{O}(1)$:

$$\frac{\int_M \exp(-\frac{1}{2t} d(x,y)^2) dV_g(x)}{\int_M \exp(-\frac{1}{2T} d(x,y)^2 - 1) dV_g(x)} \leq \frac{e^2}{1 - \exp(-\frac{1}{2})} (\frac{d+1}{d-1})^{\frac{d}{2}} \leq 20 (\frac{d+1}{d-1})^{\frac{d}{2}}.$$

Observe that $(\frac{d+1}{d-1})^{\frac{d}{2}} = (1 + \frac{1}{(d-1)/2})^{(\frac{d-1}{2} + \frac{1}{2})} = \mathcal{O}(1)$.

□