SHARP CONVERGENCE AND SAMPLING TRADE-OFFS FOR RIEMANNIAN DIFFUSION UNDER NONNEGATIVE RICCI CURVATURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models have emerged as state-of-the-art generative models, with recent extensions from Euclidean spaces to Riemannian manifolds. However, existing convergence guarantees for Riemannian diffusion models typically require $\tilde{O}(\mathrm{poly}(d,T)/\epsilon^2)$ score evaluations, with potentially unfavorable dependence on the dimension. In this work, we develop a general framework that separates score discretization from Brownian-motion simulation and allows multiple geodesic random-walk steps per score evaluation. Under nonnegative Ricci curvature assumption and an exact Brownian-motion simulation oracle, we show that $\tilde{O}(d/\epsilon^2)$ score evaluations suffice to achieve an ϵ^2 KL divergence from the target distribution, matching the existing convergence rate of Euclidean diffusion models. We further show that $\tilde{O}(d^4T/\epsilon^2)$ geodesic random-walk steps suffice to approximate the required drifted Brownian motion to ϵ total variation error. Combining these results yields a sampling scheme with $\tilde{O}(d/\epsilon^2)$ score evaluations and $\tilde{O}(d^4T/\epsilon^2)$ geodesic random-walk steps, motivating multiple random-walk steps between consecutive score evaluations. Our results provide a sharper characterization of the convergence and sampling complexity of Riemannian diffusion models.

1 INTRODUCTION

Diffusion models (Sohl-Dickstein et al., 2015) have emerged as some of the most successful generative models across a wide range of applications, including image and video generation. By modeling the generation process as stochastic differential equations (SDEs) (Song et al., 2021), diffusion model can recover data from pure noise. In particular, the gradient of the log-density, commonly referred to as the *score function*, naturally arises in the generation-time dynamics.

Most existing applications of diffusion models focus on Euclidean data. However, in many scientific and engineering applications, such as protein modeling and robotics, the underlying data have non-Euclidean structure that can be naturally represented by Riemannian manifolds. This has motivated several works to extend diffusion models from Euclidean spaces to Riemannian manifolds (De Bortoli et al., 2022; Huang et al., 2022; Lou et al., 2023). In this setting, the generation are modeled by SDEs on manifolds, with the diffusion component given by Riemannian Brownian motion.

For diffusion models in both Euclidean spaces and Riemannian manifolds, practical sampling requires discretizing the reverse-time process, and the algorithms typically require many discretization steps to achieve a small sampling error. The convergence theory of sampling algorithms for Euclidean diffusion models has been extensively studied (Chen et al., 2023b;a; Benton et al., 2024; Li & Yan, 2025). In contrast, the convergence theory for Riemannian diffusion models remains comparatively underdeveloped. Existing convergence bounds suffer from unfavorable dependence on the dimension d (Xu et al., 2026), motivating a sharper analysis of the score discretization error.

Moreover, many scientific datasets are supported on manifolds with additional geometric structure, among which *nonnegative Ricci curvature* turns to be particularly important. Examples include climate and earth events modeled on the sphere $\mathbb{S}^2$ (De Bortoli et al., 2022; Huang et al., 2022), orientations and rotations in robotics modeled on spheres $\mathbb{S}^d$ and special orthonormal groups $\mathrm{SO}(d)$ (Braun et al., 2024), and torsion angles in RNA and protein sequences modeled on flat tori $\mathbb{T}^d =$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

$(\mathbb{S}^1)^d$ (Lou et al., 2023; Cheng et al., 2025). Although these manifolds have been widely used in generation experiments, existing theoretical analyses typically impose only relatively general geometric assumptions, leaving a substantial gap between the theoretical guarantees and the settings encountered in practice.

More importantly, unlike in Euclidean space, Riemannian Brownian motion generally cannot be simulated exactly, introducing an additional source of approximation error. Existing algorithms (De Bortoli et al., 2022; Xu et al., 2026) typically perform one Brownian-motion simulation step for each score evaluation. However, as we show in our analysis, the error arising from Brownian-motion simulation may scale differently from the score discretization error. This suggests that restricting the algorithm to a single Brownian-motion simulation step per score evaluation may not provide the most efficient allocation of computational resources. In particular, when score evaluations are substantially more expensive than Brownian-motion simulation steps, it may be preferable to perform multiple Brownian-motion simulation steps between consecutive score evaluations.

In this paper, we develop a general framework for analyzing Riemannian diffusion samplers that separates the errors arising from score discretization and Brownian-motion simulation. This framework allows us to analyze multiple Brownian-motion simulation steps per score evaluation and thereby characterize the trade-off between these two sources of error. Under nonnegative Ricci curvature assumption, we obtain a sharper bound on the number of score evaluations required for accurate sampling and provide an explicit analysis of approximating Riemannian Brownian motion using geodesic random walks. Our results reveal that the number of expensive score evaluations can be substantially reduced by allocating additional, comparatively inexpensive computation to Brownian-motion simulation, providing theoretical guidance for accelerating sampling.

Our Contributions. We summarize our contributions as follows.

- We develop a general sampling framework (Section 3.3) that is sufficiently flexible to accommodate a broader class of sampling algorithms. The framework explicitly decomposes the sampling error into two components: i) the error arising from discretizing the reverse SDE, and ii) the error arising from approximating Brownian motion.
- Under our general framework, we show that, when exact simulation of Riemannian Brownian motion is available, $O\left(\epsilon^{-2} d \log^2(T/\delta)\right)$ score evaluations suffice to achieve an ϵ^2 KL divergence from the target distribution on Riemannian manifolds with nonnegative Ricci curvature (Theorem 4.3). This result substantially improves the existing convergence guarantees for Riemannian diffusion models. Moreover, it matches, up to logarithmic factors, the standard $\tilde{O}(d/\epsilon^2)$ complexity when degenerating to the Euclidean case. In addition, we also improves the existing convergence rate without Ricci curvature assumption.
- We further analyze the approximation of Riemannian Brownian motion using geodesic random walks. We show that $O\left(\epsilon^{-2} d^4 T \log(T/\delta)\right)$ geodesic random-walk steps suffice to achieve an ϵ total variation error (Theorem 5.3).
- Finally, we combine the complexities for reverse-SDE discretization and Brownian-motion simulation to characterize the trade-off between score evaluations and Brownian-motion simulation steps. In particular, our analysis yields a theoretically preferable strategy that performs multiple Brownian-motion simulation steps per score evaluation (Corollary 5.4).

2 Related Works

Convergence Analysis for Euclidean Diffusion Models. Chen et al. (2023b) developed a standard Girsanov-based framework for analyzing the sampling complexity of SDE-based generative models, including DDPMs (Ho et al., 2020). Their analysis was subsequently refined by Chen et al. (2023a); Benton et al. (2024), who established sharper KL convergence bounds under relatively mild assumptions on the score functions and data distributions. In particular, Benton et al. (2024) showed that $\tilde{O}(\epsilon^{-2} d)$ score evaluations suffice to achieve an ϵ^2-KL error. For convergence in total variation (TV) distance, early works obtained guarantees by combining KL bounds with Pinsker's inequality. More recently, Li & Yan (2025) developed a sharper analysis that reduces the complexity of achieving an ϵ-TV error to $\tilde{O}(\epsilon^{-1} d)$. In parallel, several works have focused on improving the sampling algorithms themselves to further reduce the computational complexity (Li et al., 2024b;a; Liang et al., 2024; Jain & Zhang, 2026).

	Space	Target Error	# Score Evaluations	# GRW Steps
Chen et al. (2023a)	Euclidean	$\mathrm{KL} \le \epsilon^2$	$\tilde{O}(\epsilon^{-2}d^2)$	-
Benton et al. (2024)	Euclidean	$\mathrm{KL} \le \epsilon^2$	$\tilde{O}(\epsilon^{-2}d)$	-
Li & Yan (2025)	Euclidean	$\mathrm{TV} \le \epsilon$	$\tilde{O}(\epsilon^{-1}d)$	-
Xu et al. (2026)	Manifold	$\mathrm{TV} \le \epsilon$	$\tilde{O}\left(\frac{\mathrm{poly}(\delta^{-1},T,d)}{\epsilon^2}\right)$ *	$\tilde{O}\left(\frac{\mathrm{poly}(\delta^{-1},T,d)}{\epsilon^2}\right)$
Our Results (Ric ≥ 0)	Manifold	$\mathrm{TV} \le \epsilon$	$\tilde{O}\left(\epsilon^{-2}d\right)$	$\tilde{O}\left(\epsilon^{-2}d^4T\right)$
Our Results (General)			$\tilde{O}\left(\epsilon^{-2}d^2T^2\right)$	

Table 1: Comparison with previous analyses for diffusion models.

*This complexity bound is at least $\tilde{O}(\epsilon^{-2}d^6T^2\delta^{-3})$, which comes from their Lemma 2. They did not show their explicit polynomial factor in the Brownian motion simulation error.

Riemannian Diffusion Models. De Bortoli et al. (2022) extended diffusion models from Euclidean spaces to Riemannian manifolds by formulating the forward and reverse processes as stochastic differential equations on manifolds. Independently, Huang et al. (2022) developed a related framework from a maximum-likelihood perspective. Lou et al. (2023) further improved the computation and sampling of heat kernels on Riemannian symmetric spaces, enabling Riemannian diffusion models to be applied to high-dimensional real-world tasks.

Several works have studied the design and analysis of sampling methods on Riemannian manifolds (Cheng et al., 2022; Guan et al., 2025). Regarding the convergence theory of Riemannian diffusion models, De Bortoli et al. (2022) established Wasserstein error bounds for their sampling method. More recently, Xu et al. (2026) established the first polynomial convergence rate for Riemannian diffusion models. However, their complexity bound has a high-degree polynomial dependence on the dimension d, leaving substantial room for improvement. In this work, we provide a sharper convergence analysis and show that, under nonnegative Ricci curvature, the number of score evaluations required for sampling can be reduced to $O\left(\epsilon^{-2}d\log^2(T/\delta)\right)$, which matches the standard $\tilde{O}(d/\epsilon^2)$ score-evaluation complexity of Euclidean diffusion models up to logarithmic factors.

3 Problem Formulation: Riemannian Diffusion Model

In this section, we briefly introduce the Riemannian diffusion model proposed by De Bortoli et al. (2022), and then formally propose the general sampling framework we are going to analyze.

3.1 Diffusion Model in Euclidean Spaces.

We first briefly recall diffusion models in Euclidean spaces (Song et al., 2021; Ho et al., 2020). Let p_0 be a data distribution on $\mathbb{R}^d$. We consider the Ornstein-Uhlenbeck noising process

$$\mathrm{d}X_t = -X_t\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad X_0 \sim p_0,$$

where B_t is a standard d-dimensional Brownian motion. Denote the marginal density of X_t by p_t, then p_t converges to the standard Gaussian distribution $\mathcal{N}(0, I_d)$. Fix a terminal time $T > 0$, the time-reversed process $Y_t = X_{T-t}$ satisfies

$$\mathrm{d}Y_t = (Y_t + 2\nabla \log p_{T-t}(Y_t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t. \tag{1}$$

Thus, sampling from p_0 can be performed by simulating SDE (1), initialized from $\mathcal{N}(0, I_d)$.

In practice, the score $\nabla \log p_{T-t}(\cdot)$ is approximated by a neural network $\hat{s}_t(\cdot)$. Given a time discretization $0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = T - \delta$,where $\delta > 0$ is an early-stopping time (see,

e.g., Song & Ermon 2020), Euler-Maruyama discretization of (1) leads to

$$\hat{Y}_{t_{k+1}} = \hat{Y}_{t_k} + (t_{k+1} - t_k)\left[\hat{Y}_{t_k} + 2\hat{s}_{t_k}(\hat{Y}_{t_k})\right] + \sqrt{2(t_{k+1} - t_k)}w_k, \quad w_k \sim \mathcal{N}(0, I_d).$$

Equivalently, the discretized process has a frozen drift on each interval $[t_k, t_{k+1})$:

$$\mathrm{d}\hat{Y}_t = (\hat{Y}_{t_k} + 2\hat{s}_{t_k}(\hat{Y}_{t_k}))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad t \in [t_k, t_{k+1}), \quad k = 0, \ldots, N-1. \tag{2}$$

The convergence of such sampling schemes has been extensively studied. In particular, Benton et al. (2024) showed that, under minimal assumptions, $\tilde{O}(d/\epsilon^2)$ score evaluations suffice to ensure that the KL divergence between p_δ and the distribution of $\hat{Y}_{T-\delta}$ is at most ϵ^2. This nearly linear dependence on the dimension will serve as the benchmark for our analysis of Riemannian diffusion models.

3.2 Diffusion Models on Riemannian Manifolds

We now introduce the Riemannian generalization of diffusion models (De Bortoli et al., 2022). Throughout this paper, we restrict our attention to compact manifolds. Let $\mathcal{M}$ be a connected, d-dimensional compact Riemannian manifold with metric g. We denote by $\rho(x, y)$ the Riemannian distance between $x, y \in \mathcal{M}$, by ∇ the Riemannian gradient, and by Δ the Laplace-Beltrami operator. We use $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_g$ to denote the Riemannian inner product.

Let p_0 be the data distribution on $\mathcal{M}$. To define a diffusion process on $\mathcal{M}$, we replace the Euclidean Brownian motion by Riemannian Brownian motion. We use the following characterization, which is convenient for our analysis.

Definition 3.1 (Brownian motion on $\mathcal{M}$)**.** *Let $(X_t)_{t\geq 0}$ be a continuous stochastic process on $\mathcal{M}$, and let b_t be a time-dependent vector field on $\mathcal{M}$. We say that X_t is a Riemannian Brownian motion on $\mathcal{M}$ with drift b_t, if, for any smooth function $f : [0, \infty) \times \mathcal{M} \to \mathbb{R}$, the process M_t^f defined by*

$$M_t^f := f(t, X_t) - f(0, X_0) - \int_0^t \left(\partial_s + \langle b_s(X_s), \nabla \rangle + \frac{1}{2}\Delta\right) f(s, X_s)\mathrm{d}s$$

is a continuous local martingale. Equivalently, we say $\mathrm{d}X_t = b_t(X_t)\mathrm{d}t + \mathrm{d}B_t$.

For compact manifolds, we consider the following forward noising process:

$$\mathrm{d}X_t = \mathrm{d}B_t, \quad X_0 \sim p_0. \tag{3}$$

For $t > 0$, X_t admits a smooth density $p_t(\cdot)$ with respect to the Riemannian volume measure, satisfying the heat equation

$$\partial_t p_t(x) = \frac{1}{2}\Delta p_t(x).$$

Since $\mathcal{M}$ is compact and connected, the operator $-\Delta$ has nonnegative eigenvalues. So, p_t converges to the uniform distribution $u_{\mathcal{M}}$, given by the normalized Riemannian volume measure.

For a terminal time $T > 0$, define the reverse process by $Y_t = X_{T-t}$. Under standard regularity conditions, the reverse process satisfies (De Bortoli et al., 2022)

$$\mathrm{d}Y_t = \nabla \log p_{T-t}(Y_t)\mathrm{d}t + \mathrm{d}B_t. \tag{4}$$

Thus, given access to the Riemannian score $\nabla \log p_{T-t}(\cdot)$ and an initial sample approximately distributed according to $u_{\mathcal{M}}$, sampling from p_0 reduces to simulating the reverse SDE.

In practice, the score is approximated by a neural network $\hat{s}_t(\cdot) \approx \nabla \log p_{T-t}(\cdot)$. A common approach to approximate the reverse SDE is based on geodesic random walks, which can be viewed as the Riemannian analogue of Euler-Maruyama. Let $0 = t_0 < t_1 < \cdots < t_N = T - \delta$ be the time discretization, and denote the step sizes by $h_k := t_{k+1} - t_k$. Starting from $\hat{Y}_0$, the geodesic random walk considered in De Bortoli et al. (2022) is given by

$$\hat{Y}_{t_{k+1}} = \exp_{\hat{Y}_{t_k}}\left(h_k \hat{s}_{t_k}(\hat{Y}_{t_k}) + \sqrt{h_k} w_k\right), \quad k = 0, \ldots, N-1, \tag{5}$$

where $\exp_x \mathcal{M} \to \mathcal{M}$ is the exponential map and w_k is a standard Gaussian vector in $T_{\hat{Y}_{t_k}}\mathcal{M}$ with respect to any orthonormal basis. Equivalently, each step first performs an Euler-Maruyama update in the tangent space and then maps the resulting tangent vector to $\mathcal{M}$ through the exponential map.

Unlike in Euclidean space, however, Riemannian Brownian motion generally does not admit an exact finite-step simulation by a Gaussian increment. Therefore, the error of (5) has two distinct sources: the error from discretizing the reverse SDE and the error from approximating Riemannian Brownian motion. The latter arises from the local approximation $\sqrt{h_k}w_k$ on the tangent space, which does not exactly reproduce a Riemannian Brownian increment over a finite time interval. Existing convergence analyses typically treat these two sources of error together, resulting in bounds with unfavorable dependence on the dimension d. For example, Xu et al. (2026) showed that, under suitable assumptions on the manifold and score functions, we require $O\left(\mathrm{poly}(d, T, \delta^{-1})/\epsilon^2\right)$ steps to ensure that the total variation (TV) distance between the distributions of X_δ and $\hat{Y}_{T-\delta}$ is at most ϵ.

3.3 A Sampling Framework: Brownian Motion Simulation Oracle

As discussed in the previous section, sampling from the reverse SDE on a Riemannian manifold involves two distinct difficulties: discretizing the reverse SDE, and simulating Riemannian Brownian motion. To separate these two sources of difficulty, we introduce a general sampling framework based on a Brownian-motion simulation oracle (i.e., a black-box).

Consider a time discretization $0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = T - \delta$, and let $h_k := t_{k+1} - t_k$ denote the length of the k-th time interval. We assume access to a general Brownian-motion simulation oracle $\mathcal{O}$. Given a time-independent vector field b on $\mathcal{M}$, an initial point $x \in \mathcal{M}$, and a time horizon $h > 0$, the oracle $\mathcal{O}(b, x, h)$ returns a random point whose distribution approximates the endpoint at time h of the Riemannian SDE

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \mathrm{d}B_t, \quad X_0 = x, \quad t \in [0, h]. \tag{6}$$

The oracle is allowed to be implemented using an arbitrary simulation procedure; in particular, it need not be restricted to a single geodesic random-walk step.

The high-level idea of our framework is simple. At each step k, we first evaluate the score estimator at the current state to obtain an instantaneous drift s_k. We then extend this drift to a vector field $\tilde{s}_k$ on $\mathcal{M}$, and use the Brownian-motion simulation oracle to evolve the process over the interval $[t_k, t_{k+1}]$. Formally, given the initial state $\hat{Y}_0$, our sampling framework is summarized in Algorithm 1.

Algorithm 1 Sampling Framework with Simulation Oracle.

Require: $\{t_k\}_{k=0}^N$, $\mathcal{O}(\cdot,\cdot,\cdot)$, $\hat{s}_{t_k}(\cdot)$, $\hat{Y}_0$.
1: **for** $k = 0, 1, \ldots, N-1$ **do**
2: $\quad s_k \leftarrow \hat{s}_{t_k}(\hat{Y}_{t_k})$, $h_k \leftarrow t_{k+1} - t_k$.
3: $\quad$ Define a vector field $\tilde{s}_k(\cdot)$ using $\hat{Y}_{t_k}$ and s_k, satisfying $\tilde{s}(\hat{Y}_{t_k}) = s_k$.
4: $\quad \hat{Y}_{t_{k+1}} \leftarrow \mathcal{O}(\tilde{s}_k, \hat{Y}_{t_k}, h_k)$.
5: **end for**

Suppose first that the oracle provides an exact simulation of SDE (6). Then the resulting sampling process can be represented by the SDE

$$\mathrm{d}\hat{Y}_t = \tilde{s}_k(\hat{Y}_t)\mathrm{d}t + \mathrm{d}B_t, \quad t \in [t_k, t_{k+1}). \tag{7}$$

Thus, when the Brownian-motion simulation is exact, the only approximation relative to the true reverse process (4) comes from replacing the time-dependent score $\nabla \log p_t(\cdot)$ by the frozen vector field $\tilde{s}_k(\cdot)$ on each time interval. Consequently, the distributional error between $\hat{Y}_{T-\delta}$ and the target $X_\delta = Y_{T-\delta}$ can be analyzed entirely through a comparison between the two SDEs (4) and (7).

This formulation is closely analogous to Euler-Maruyama discretization in Euclidean space. The key difference is that, in $\mathbb{R}^d$, an instantaneous drift s_k is naturally identified with a constant vector field, so one can simply set $\tilde{s}_k(x) = s_k$. On a Riemannian manifold, however, $s_k \in T_{\hat{Y}_{t_k}}\mathcal{M}$ belongs to a tangent space that depends on the current point, and therefore cannot be directly evaluated at other points on $\mathcal{M}$. Constructing a suitable extension $\tilde{s}_k$ is therefore an additional geometric component of the discretization. We provide an appropriate construction and analyze its resulting error in Section 4.

For a general, potentially inexact simulation oracle, the total distributional error can be decomposed into two components. The first is the error incurred by the frozen-drift approximation assuming

that the Brownian-motion simulation is exact; the second is the additional error introduced by the simulation oracle itself. This decomposition allows us to analyze score discretization and Brownian-motion simulation separately. A similar decomposition has also been used implicitly in the analysis of Xu et al. (2026), but our framework makes the two sources of error explicit and allows the Brownian-motion simulation procedure to be designed independently of the score discretization.

As a special case, suppose that the oracle is defined by $\mathcal{O}(b, x, h) = \exp_x \left(h \cdot b(x) + \sqrt{h} w \right)$ where $w \in T_x\mathcal{M}$ is an independent Gaussian vector. Then Algorithm 1 reduces exactly to the geodesic random walk introduced in the previous section, regardless of how the vector field $\tilde{s}_k(\cdot)$ is defined away from $\hat{Y}_{t_k}$. More generally, however, our framework permits arbitrarily sophisticated Brownian-motion simulation procedures while requiring only a single score evaluation per interval. This distinction is important because score evaluations, typically implemented by evaluating a neural network, can be substantially more expensive than Brownian-motion simulation steps. In particular, in Section 5, we instantiate the oracle $\mathcal{O}$ by performing n_{bm} smaller geodesic random-walk steps within each interval $[t_k, t_{k+1}]$. Thus, we can increase the accuracy of Brownian-motion simulation without increasing the number of score evaluations. This provides the basic mechanism underlying our computational trade-off between score evaluations and Brownian-motion simulation steps.

In the following sections, we first analyze the idealized setting in which the Brownian-motion simulation oracle is exact. Under nonnegative Ricci curvature, we show in Section 4 that nearly linear dependence on the dimension d is sufficient for the number of score evaluations. We then analyze the additional error introduced by approximating the oracle using n_{bm} geodesic random-walk steps in Section 5.

4 Sharp Convergence Rate with Exact Simulation Oracle under Nonnegative Ricci Curvature

In this section, we analyze Algorithm 1 on manifolds with nonnegative Ricci curvature under the assumption that the simulation oracle $\mathcal{O}$ exactly solves the Riemannian SDE (6). Our goal is therefore to compare the true reverse process (4) with the estimated process (7). We adopt the construction of Xu et al. (2026). Let $\eta : \mathbb{R} \to \mathbb{R}$ be a C^2 cut-off function, satisfying

$$\eta(r) = \begin{cases} 1, & |r| \le \frac{1}{2}, \\ 0, & |r| \ge 1, \end{cases} \quad |\eta'(r)| \le C, \quad |\eta"(r)| \le C,$$

where $C > 0$ is a universal constant. For a scale parameter $\omega > 0$, we define $\eta_\omega : r \mapsto \eta(r/\omega)$. Thus, η_ω is supported on $[-\omega, \omega]$. Fix a discretization time t_k, let $y := \hat{Y}_{t_k} = y$, $s_k := \hat{s}_{t_k}(y)$. With these information, we then define the vector field in Algorithm 1 by

$$\tilde{s}_k(x) := \eta_\omega(\rho(x, y)) \cdot \mathrm{d}[\exp_y \left(\log_y x \right)](s_k), \quad x \in \mathcal{M}. \tag{8}$$

where $\log_y x$ denotes the inverse exponential map on the normal neighborhood of y. Here, we canonically identify $T_{\log_y x}(T_y\mathcal{M})$ with $T_y\mathcal{M}$, since $T_y\mathcal{M}$ is a vector space. The construction has a simple geometric interpretation. In normal coordinates centered at y, the vector s_k is kept constant, and is then mapped back to the manifold through the differential of the exponential map. The cut-off function localizes this construction to the geodesic ball $B_y(\omega)$, while also ensuring that $\tilde{s}_k$ is globally well-defined and smooth. In particular, $\tilde{s}_k(y) = s_k$, and within the smaller ball $B_y(\omega/2)$, the coefficients of $\tilde{s}_k$ in the normal coordinate representation $x \mapsto \log_y x$ are constant.

We take the scale parameter as $\omega := c/\sqrt{d}$ in the construction (8) for a sufficiently small universal constant $c > 0$. Under the geometric assumptions imposed later, this choice can ensure that the relevant normal neighborhood is contained within the injectivity radius, and the Riemannian metric g is sufficiently close to the identity matrix in the local coordinate representation. For manifolds with an explicit exponential map, such as the sphere, the vector field (8) can be computed explicitly.

Now, we introduce the assumptions required for our theory.

Assumption 4.1. *The compact manifold $\mathcal{M}$ satisfies the following geometric conditions.*

- ***Bounded diameter.*** $\mathrm{diam}(\mathcal{M}) = \sup\{\rho(x, y) : x, y \in \mathcal{M}\} \le D$.

- ***Bounded injectivity radius.*** *The injective radius $i(x)$ satisfies $i(x) \geq r_i > 0$.*
- ***Bounded curvature.*** *The Riemannian curvature tensor R satisfies $\max\{\|R\|, \|\nabla R\|, \|\nabla^2 R\|\} \leq C_{\mathrm{Rm}}$, where the norms are operator norms.*
- ***Nonnegative Ricci curvature.*** *For every $x \in \mathcal{M}$ and $X \in T_x\mathcal{M}$, $\mathrm{Ric}(X, X) \geq 0$.*

The first three conditions in Assumption 4.1 are the same geometric assumptions considered by Xu et al. (2026). For any fixed compact smooth Riemannian manifold, these quantities are finite and the injectivity radius is strictly positive. In our analysis, we additionally require the corresponding constants D, r_i, and C_{Rm} to be independent of the dimension d, since we are interested in the dependence of the convergence rate on d. This is satisfied, for example, on the unit d-sphere $\mathbb{S}^d$, where we can take $D = \pi, r_i = \pi$ and $C_{\mathrm{Rm}} = 1$, which are all independent of d.

The **nonnegative Ricci curvature** assumption is the main additional geometric condition in our analysis. It does not hold for general compact Riemannian manifolds; nevertheless, it is satisfied by many commonly used manifolds, including the unit sphere $\mathbb{S}^d$ and the special orthogonal group $\mathrm{SO}(n)$ equipped with its standard bi-invariant metric. We discuss how the analysis can be extended to manifolds with general Ricci curvature in Appendix C. As we show later, nonnegative Ricci curvature plays a crucial role in obtaining the nearly d-linear KL convergence bound.

Assumption 4.2 (Score Error)**.** $\sum_{k=0}^{N-1} h_k \mathbb{E}_{x\sim p_{T-t_k}} \left[\|\nabla \log p_{T-t_k}(x) - \hat{s}_{t_k}(x)\|^2\right] \leq \epsilon^2_{\mathrm{score}}$.

Assumption 4.2 is standard in the theoretical analysis of diffusion models (Chen et al., 2023b) and measures the cumulative mean-squared error of the score estimator along the forward diffusion process. In particular, our analysis only requires an error bound in expectation and does not impose a uniform pointwise bound on the score estimation error.

With Assumptions 4.1 and 4.2 in place, we are ready to state our main convergence result. Recall that (p_t) denotes the marginal density of the forward process. Let $(\hat{q}_t)$ denote the marginal density of estimated reverse process $(\hat{Y}_t)_{t\geq 0}$ defined by SDE (7).

Theorem 4.3. *Fix $T > 0$ and $\delta > 0$. Let the step sizes be given by $h_k = \min\{h(T - t_k), T - \delta - t_k\}$, where $0 < h < \frac{1}{2}$ is a scale parameter. Suppose that the vector field used in Algorithm 1 is defined according to (8). Then, under Assumptions 4.1 and 4.2,*

$$\begin{aligned}\mathrm{TV}(p_\delta, \hat{q}_{T-\delta})^2 &\lesssim \mathrm{KL}(p_\delta \| \hat{q}_{T-\delta}) \\ &\lesssim \underbrace{\mathrm{KL}(p_T \| \hat{q}_0)}_{\text{Initialization error}} + \epsilon^2_{score} + \underbrace{hd\log(T/\delta) + h^2 d^4 T^2 + Ch^3 d^7 T^3}_{\text{Discretization error}}.\end{aligned}$$

Here, the constants hidden by $\lesssim$ (i.e., $= O(\cdot)$) are independent of the manifold dimension d, but may depend on the geometric constants in Assumption 4.1.

Theorem 4.3 decomposes the distributional error into three components: i) the *initialization error*, which arises from the discrepancy between the initial distribution of the sampling process and the desired initial distribution of the reverse process; ii) the *score estimation error*, which quantifies the accuracy of the estimated score; and iii) the *discretization error*, which arises from freezing the vector field over each time interval $[t_k, t_{k+1}]$. The first two terms do not depend on the step-size parameter h, and indicates the error we can achieve as the discretization becomes arbitrarily fine. In contrast, the discretization error vanishes as $h \to 0$, and its dependence on h determines the convergence rate of the sampling scheme. We show the full proof of Theorem 4.3 in Appendix B.

If we use the step sizes specified in Theorem 4.3, the total number of discretization steps satisfies $N = \Theta\left(h^{-1}\log(T/\delta)\right)$. Suppose further that $\mathrm{KL}(p_T\|\hat{q}_0) + \epsilon^2_{\mathrm{score}} = O(\epsilon^2)$. Then Theorem 4.3 implies that, to achieve an $O(\epsilon^2)$ KL error (and hence an $O(\epsilon)$ TV error) for sufficiently small $\epsilon > 0$, it suffices to choose

$$N = O\left((\epsilon^{-2} d\log(T/\delta) + \epsilon^{-1} d^2 T + \epsilon^{-2/3} d^{7/3} T)\log(T/\delta)\right) = O\left(\epsilon^{-2} d\log^2(T/\delta)\right).$$

For comparison, the analysis of Xu et al. (2026) yields $N = O\left(\frac{d^6 T^2}{\delta^3 \epsilon^2}\right)$ in the setting considered here[1]. Our result substantially improves the dependence on the dimension and replaces the polynomial

[1]We obtain this corresponding step-size bound from Lemma 2 of Xu et al. (2026), which applies precisely to the setting considered in this section.

dependence on $1/\delta$ with a polylogarithmic dependence. Moreover, up to logarithmic factors, the number of discretization steps has the same dependence on d and ϵ as the as the convergence guarantee for diffusion models in Euclidean space established by Benton et al. (2024). In particular, for compact manifolds with nonnegative Ricci curvature, our analysis shows that the dimension dependence can be reduced to essentially linear. An overview of the argument leading to this nearly linear dependence on d is provided in Appendix A.

Remark on General Compact Manifolds. Our analysis for Theorem 4.3 can be applied without the nonnegative Ricci curvature assumption. We provide the discussions in Appendix C, where we modify Assumption 4.1 so that the Ricci curvature is instead bounded below by $-C_R$ for some constant $C_R > 0$. This assumption always holds for compact manifolds, and the definition of Ricci curvature directly implies that $C_R \le C_{\mathrm{Rm}} d$. In the worst case $C_R = C_{\mathrm{Rm}} d$, our analysis still implies a bound of $N = \tilde{O}\left(\frac{d^2 T^2}{\epsilon^2}\right)$ on the number of discretization steps, which improves the bound $O\left(\frac{d^6 T^2}{\delta^3 \epsilon^2}\right)$ provided in (Xu et al., 2026) by d^4 in dimension for general compact smooth manifolds.

5 Convergence with Geodesic Random Walk

In the previous section, we analyzed the distributional error between the true reverse process (4) and the approximated reverse process (7) under an exact Brownian-motion simulation oracle. We now analyze the additional error introduced by implementing the oracle using geodesic random walks. Specifically, we perform n_{bm} geodesic random walk steps within each interval of Algorithm 1. The procedure is given in Algorithm 2. When $n_{\mathrm{bm}} = 1$, this reduces to the standard geodesic random-walk scheme. Increasing n_{bm} improves the accuracy of Brownian-motion simulation without increasing the number of score evaluations. As we show below, its convergence rate differs from that of the score discretization error analyzed in the previous section.

Algorithm 2 Implementing $\mathcal{O}(b, x, h)$ using n_{bm} geodesic random walks.

Require: b, $x \in \mathcal{M}$, $h \ge 0$, $n_{\mathrm{bm}} \in \mathbb{Z}_+$.
1: $h' \leftarrow h/n_{\mathrm{bm}}$.
2: $x_0 \leftarrow x$.
3: **for** $k = 0, 1, \ldots, n_{\mathrm{bm}} - 1$ **do**
4: Sample Gaussian random vector w_k in $T_{x_k}\mathcal{M}$.
5: $v_k \leftarrow h' \cdot b(x_k) + \sqrt{h'} w_k$.
6: $x_{k+1} \leftarrow \exp_{x_k}(v_k)$.
7: **end for**
8: $\mathcal{O}(b, x, h) \leftarrow x_{n_{\mathrm{bm}}}$.

To analyze this approximation, we require the following additional regularity assumptions.

Assumption 5.1. *In addition to Assumption 4.1, assume that* $\max\{\|\nabla^3 R\|, \|\nabla^4 R\|\} \le C'_{\mathrm{Rm}}$.

Similarly as in Assumption 4.1, we assume that C'_{Rm} is independent of the dimension d. This additional regularity is needed to control higher-order derivatives of the metric tensor g.

Assumption 5.2. *For every* $k = 0, \ldots, N-1$ *and* $x \in \mathcal{M}$, $\|\hat{s}_{t_k}(x)\| \le C_{\mathrm{score}}\left(\frac{D}{T - t_k} + \sqrt{\frac{d}{T - t_k}}\right)$, *where* C_{score} *is independent of* d, *and* D *is the diameter of* $\mathcal{M}$.

Assumption 5.2 imposes a uniform bound on the estimated score. By Lemma E.2, the true score under nonnegative Ricci curvature satisfies the same bound. In this case, we may clip $\hat{s}_{t_k}$ to this range without increasing its pointwise estimation error, so Assumption 4.2 will remain valid after clipping.

We are now ready to state the convergence guarantee for the geodesic random walk approximation. In Algorithm 1, we continue to use (8) to define the velocity field $\tilde{s}_k(\cdot)$. Let $\hat{Y}_{T-\delta}$ denote the solution to SDE (7), and let $\tilde{Y}_{T-\delta}$ denote the output of Algorithm 1 when the Brownian-motion simulation oracle is implemented by Algorithm 2. We assume that both processes are initialized from the same distribution, $\hat{Y}_0 \sim \hat{q}_0$. The following theorem bounds the additional error introduced by the geodesic random-walk simulation.

Theorem 5.3. *Let $\hat{q}_{T-\delta}$ and $\tilde{q}_{T-\delta}$ denote the distributions of $\hat{Y}_{T-\delta}$ and $\tilde{Y}_{T-\delta}$, respectively. Let $h' := h/n_{bm}$ denote the inner step size scale, where $h > 0$ is the scale parameter defined in Theorem 4.3. Under Assumptions 4.1, 5.1 and 5.2, if h' is sufficiently small, then*

$$\mathrm{TV}(\hat{q}_{T-\delta}, \tilde{q}_{T-\delta})^2 \lesssim h' \left(d^4 T + \delta^{-2} D^4 + D^2 d^3 \log(T/\delta) \right),$$

where the constants hidden by $\lesssim$ are independent of the dimension d.

The detailed proof of Theorem 5.3 is provided in Appendix D. Although we use a parametrix-based analysis similar to that of Xu et al. (2026), our refined estimates lead to a substantially improved dependence on the dimension d.

Theorem 5.3, shows that the squared TV error scales linearly with the inner step size h', as in the discretization error in Theorem 4.3. However, its dependence on the dimension is substantially worse. This indicates that accurately simulating drifted Brownian motion may require more computational steps than discretizing the score function alone. Combining the two error bounds yields the following result, whose proof is given inD.6.

Corollary 5.4. *Suppose Assumptions 4.1, 4.2, 5.1 and 5.2 hold. Assume further that $\delta = \Omega(d^{-2})$, $\mathrm{TV}(\hat{q}_0, p_T) + \epsilon_{score} = O(\epsilon)$. Then, for sufficiently small $\epsilon > 0$, under the step-size schedule of Theorem 4.3, it suffices to choose*

$$N = O\left(\epsilon^{-2} d \log^2(T/\delta)\right), \quad N' = O\left(\epsilon^{-2} d^4 T \log(T/\delta)\right)$$

to ensure $\mathrm{TV}(p_\delta, \tilde{q}_{T-\delta}) = O(\epsilon)$. Here, N is the number of evaluations of score function, and $N' = N n_{bm}$ is the total number of geodesic random-walk steps.

Corollary 5.4 provides complexity bounds for both score evaluations and geodesic random-walk steps. Since evaluating the neural-network score is typically substantially more expensive than simulating a single geodesic random-walk step, it can be computationally advantageous to perform multiple random-walk steps between consecutive score evaluations. Thus, the corollary provides a theoretical justification for using multiple Brownian-motion simulation steps per score evaluation.

Remark on the time horizon T. The initial distribution $\hat{q}_0$ is typically chosen to be the uniform distribution $u_{\mathcal{M}}$ with respect to the Riemannian volume measure. Therefore, T must be sufficiently large to ensure that that $\mathrm{TV}(p_T, u_{\mathcal{M}}) = O(\epsilon)$. As shown by (Xu et al., 2026), it suffices to choose $T = O\left(\lambda_1^{-1}(d \log d + \log(1/\epsilon))\right)$, where λ_1 is the smallest positive eigenvalue of $-\Delta$. For general manifold, obtaining an explicit lower bound on λ_1 can be difficult. However, for the d-dimensional unit sphere, it is known that $\lambda_1 = d$, giving $T = O(\log d + d^{-1} \log(1/\epsilon))$.

6 Conclusion

In this work, we developed a sharp convergence analysis for Riemannian diffusion models. We introduced a general sampling framework (Section 3.3) that explicitly separates the sampling error into score discretization error and Brownian-motion simulation error. Under an exact Brownian-motion simulation oracle, our analysis shows that $\tilde{O}(\epsilon^{-2} d)$ score evaluations suffice to achieve an ϵ^2 KL error (Theorem 4.3), substantially improving existing convergence guarantees and matching the nearly linear dimension dependence known for Euclidean diffusion models. We further analyzed the Brownian-motion simulation error of geodesic random walks and showed that, from the resulting error trade-off, performing multiple geodesic random-walk steps per score evaluation can be computationally advantageous.

Several directions remain open for future work. First, our sampling framework is sufficiently general to accommodate more sophisticated sampling schemes, leaving room for further improvements through better choices of the vector-field extension or more accurate and efficient Brownian-motion simulation methods. Second, our TV guarantees are obtained via Pinsker's inequality from the KL bounds. Analyses of Euclidean diffusion models suggest that a direct analysis of TV error may yield sharper convergence rates, and it would be interesting to investigate whether similar improvements are possible on Riemannian manifolds. Finally, our analysis focuses on the SDE formulation of diffusion sampling, while the convergence properties of the corresponding probability-flow ODE on Riemannian manifolds remain largely unexplored.

AI USE STATEMENT

In this work, we used generative AI tools to polish writing. We have not used generative AI tools for retrieval, discovery, research ideation or execution, drafting sections of the paper and proving mathematical claims. Generating synthetic datasets are not applicable to this work. We have checked the content polished by LLM to ensure accurate representation of our claims. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *International Conference on Learning Representations*, 2024.

Nicole Berline, Ezra Getzler, and Michèle Vergne. *Heat Kernels and Dirac Operators*. Springer, 2nd edition, 2004.

Max Braun, Noémie Jaquier, Leonel Rozo, and Tamim Asfour. Riemannian flow matching policy for robot motion learning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, 2023a.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023b.

Chaoran Cheng, Yusong Wang, Yuxin Chen, Xiangxin Zhou, Nanning Zheng, and Ge Liu. Riemannian consistency model. In *Advances in Neural Information Processing Systems*, 2025.

Xiang Cheng, Jingzhao Zhang, and Suvrit Sra. Efficient sampling on riemannian manifolds via langevin mcmc. In *Advances in Neural Information Processing Systems*, 2022.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2022.

A. Engoulatov. A universal bound on the gradient of logarithm of the heat kernel for manifolds with bounded ricci curvature. *Journal of Functional Analysis*, 2006.

Yunrui Guan, Krishnakumar Balasubramanian, and Shiqian Ma. Riemannian proximal sampler for high-accuracy sampling on manifolds. In *Advances in Neural Information Processing Systems*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Elton Hsu. Estimates of derivatives of the heat kernel on a compact riemannian manifold. *Proceedings of the American Mathematical Society*, 1999.

Elton P. Hsu. *Stochastic Analysis on Manifolds*. 2002.

Chin-Wei Huang, Milad Aghajohari, Joey Bose, Prakash Panangaden, and Aaron Courville. Riemannian diffusion models. In *Advances in Neural Information Processing Systems*, 2022.

Nishant Jain and Tong Zhang. A sharp kl convergence analysis for diffusion models under minimal assumptions. In *International Conference on Learning Representations*, 2026.

Jürgen Jost. *Riemannian geometry and geometric analysis*. Springer, 2005.

Gen Li and Yuling Yan. $o(d/t)$ convergence theory for diffusion probabilistic models under minimal assumptions. In *The Thirteenth International Conference on Learning Representations*, 2025.

Gen Li, Yu Huang, Timofey Efimov, Yuting Wei, Yuejie Chi, and Yuxin Chen. Accelerating convergence of score-based diffusion models, provably, 2024a.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards faster non-asymptotic convergence for diffusion-based generative models, 2024b.

Yuchen Liang, Peizhong Ju, Yingbin Liang, and Ness Shroff. Broadening target distributions for accelerated diffusion models via a novel analysis approach, 2024.

Aaron Lou, Minkai Xu, Adam Farris, and Stefano Ermon. Scaling riemannian diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*. pmlr, 2015.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Xingyu Xu, Ziyi Zhang, Yorie Nakahira, Guannan Qu, and Yuejie Chi. Polynomial convergence of riemannian diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026.

NOTATIONS

Throughout the proof, we use C to represent constants that are independent of d, and for notational simplicity, the value of C may change from line to line. We use greek alphabet (i.e., $\alpha, \beta, \eta, \gamma$, etc.) to denote the indices of tensors in local coordinates, and we assume Einstein's summation convention.

We use $\mathcal{M}$ to denote the Riemannian manifold, and g represents the Riemannian metric. We write $\sqrt{g} := \sqrt{\det(g)}$ as the volume unit. $\int_{\mathcal{M}} f(x) \cdot \sqrt{g(x)} dx$ denotes the integration on $\mathcal{M}$ with respect to the volume measure. We also simply write $\int_{\mathcal{M}} f(x) dx$ to omit the volume unit.

We use $T\mathcal{M}$ to represent the tangent bundle of $\mathcal{M}$, and $T_x\mathcal{M}$ to denote the tangent space at $x \in \mathcal{M}$. The basis for $T_x\mathcal{M}$ in the local coordinate representation is denoted by $\{\partial_\alpha\}_{\alpha=1}^d$ or $\{e_\alpha\}_{\alpha=1}^d$. For a smooth function $f(\cdot)$ on the manifold, we use ∇f to represent its Riemannian gradient, which satisfies

$$\forall x \in \mathcal{M}, \forall v \in T_x\mathcal{M}, \quad \langle \nabla f(x), v \rangle = \mathrm{d}f(x)(v).$$

We use Christoffel symbols

$$\Gamma_{\alpha\beta}^\gamma := \frac{1}{2} g^{\gamma\eta} \left(\partial_\alpha g_{\beta\eta} + \partial_\beta g_{\alpha\eta} - \partial_\eta g_{\alpha\beta} \right),$$

The Levi-civita connection is given by

$$\nabla_\alpha V (= \nabla_{e_\alpha} V) = \partial_\alpha v^\beta e_\beta + \Gamma_{\alpha\beta}^\gamma v^\beta e_\gamma,$$

where $V(\cdot) = v^\beta e_\beta$ is a vector field on $\mathcal{M}$. The Levi-civita connection can be extended for general tensors, see e.g., Chapter 4 of (Jost, 2005).

For Riemannian curvature, we use the convension

$$R(X, Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z$$

for vector fields X, Y and Z, where $[\cdot, \cdot]$ denotes the Lie bracket

$$[X, Y] = \left[X^\beta \partial_\beta Y^\alpha - Y^\beta \partial_\beta X^\alpha \right] e_\alpha, \quad X = X^\alpha e_\alpha, Y = Y^\alpha e_\alpha.$$

We use the convension

$$\|R\| \leq \sup_{\|X\|=\|Y\|=\|Z\|=1} \|R(X, Y)Z\|$$

for the operator norm, which is well-defined because the value of $R(X, Y)Z$ at a point $x \in \mathcal{M}$ only depends on the value of X, Y, Z at x. The norm is defined similarly for the derivatives. Moreover, the Ricci curvature is given by

$$\mathrm{Ric}(X, Y) = g^{\alpha\beta} \langle R(X, e_\alpha) e_\beta, Y \rangle,$$

i.e., the trace of the Riemannian curvature tensor.

The Laplace-Beltrami operator on $\mathcal{M}$ for smooth functions is defined by

$$\Delta f = \mathrm{div}(\nabla f) = \frac{1}{\sqrt{g}} \partial_\alpha \left(\sqrt{g} g^{\alpha\beta} \partial_\beta f \right) = g^{\alpha\beta} \left(\partial_{\alpha\beta} - \Gamma_{\alpha\beta}^\gamma \partial_\gamma \right) f.$$

Note that some mathematical books have also used the definition $-\Delta$, to ensure nonnegative eigenvalues. For a general tensor T, we use the extension given by

$$\Delta T = g^{\alpha\beta} \left(\nabla_\alpha \nabla_\beta - \Gamma_{\alpha\beta}^\gamma \nabla_\gamma \right) T.$$

For any linear operator $T : H_1 \to H_2$ between Hilbert spaces, we define its Hilbert-Schmidt norm by

$$\|T\|_{\text{HS}} = \sqrt{\sum_i \|Te_i\|^2}, \quad \{e_i\} \text{ is an orthonormal basis of } H_1.$$

Clearly, when H_1 and H_2 are Euclidean spaces, the Hilbert-Schimidt norm is exactly the Frobenius norm of matrix representation of T in orthonormal bases.

A Improved Analysis to Achieve d-linear Dependence

In Theorem 4.3, the dominant term for discretization error is $hd\log(T/\delta)$, which is linear in d. This eventually produces the d-linear convergence rate for the sampling process. In this section, we sketch through some key technical improvements that lead to this bound.

Our goal is to compare the generated distributions given by SDEs (4) and (7). Since both equations share the same Brownian motion part, a standard Girsanov-based approach indicates that

$$\mathrm{KL}(P\|\hat{P}) = \mathrm{KL}(p_T\|\hat{q}_0) + \frac{1}{2}\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} \mathbb{E}_{Y\sim P}\left[\|\nabla \log p_{T-t}(Y_t) - \tilde{s}_k(Y_t)\|^2\right] \mathrm{d}t,$$

where P and $\hat{P}$ are path measures of $(Y_t)_{t\geq 0}$ and $(\hat{Y}_t)_{t\geq 0}$, respectively. By extracting the score matching error and apply the Definition 3.1 of geometric Brownian motion (i.e., Itô's rule) on the score difference, where we skip the detailed calculations here, the KL divergence are eventually bounded by

$$\mathrm{KL}(P\|\hat{P}) \lesssim \mathrm{KL}(p_T\|\hat{q}_0) + \epsilon_{\text{score}}^2 + \sum_{k=0}^{N-1} h_k \int_{t_k}^{t_{k+1}} \mathbb{E}_{Y\sim P}\left[\|\nabla^2 \log p_{T-t}(Y_t)\|_{\text{HS}}^2\right] \mathrm{d}t + o(h).$$

In the above error bound, the dominating term consists of the expectation of squared Hilbert-Schmidt norm for Hessians of log density. However, until now, it is still insufficient to obtain the d-linear error bound, because even for the case of Euclidean diffusion model, the worst-case bound for $\mathbb{E}_{Y\sim P}\left[\|\nabla^2 \log p_{T-t}(Y_t)\|_{\text{HS}}^2\right]$ is still of order $O(d^2)$. To further reduce the degree of d, we prove two key inequalities under the assumption of nonnegative Ricci curvature:

$$(T-t)\mathbb{E}[\|\nabla \log p_{T-t}(Y_t)\|_{\text{HS}}^2] \leq \frac{\mathrm{d}}{\mathrm{d}t}(T-t)\mathbb{E}[\|\nabla \log p_{T-t}(Y_t)\|^2] + \mathbb{E}[\|\nabla \log p_{T-t}(Y_t)\|^2]; \quad (9)$$

$$\mathbb{E}_{x\sim p_t}\left[\|\nabla \log p_t(x)\|^2\right] \leq \frac{Cd}{t}, \quad C \text{ is a universal constant.} \quad (10)$$

Therefore, apply inequalities (9) and (10), and note that our selected step size schedule satisfies that $h_k = O(h(T-t))$ for $t \in [t_k, t_{k+1}]$,

$$\begin{aligned}
&\sum_{k=0}^{N-1} h_k \int_{t_k}^{t_{k+1}} \mathbb{E}_{Y\sim P}\left[\|\nabla^2 \log p_{T-t}(Y_t)\|_{\text{HS}}^2\right] \mathrm{d}t \\
&\leq Ch \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} (T-t)\mathbb{E}_{Y\sim P}\left[\|\nabla^2 \log p_{T-t}(Y_t)\|_{\text{HS}}^2\right] \mathrm{d}t \\
&\leq Ch\left(\delta\mathbb{E}[\|\nabla \log p_\delta(X_\delta)\|^2] + \int_\delta^T \mathbb{E}[\|\nabla \log p_t(X_t)\|^2]\mathrm{d}t\right) \\
&\leq Ch\left(d + \int_\delta^T \frac{d}{t}\mathrm{d}t\right) = O\left(hd\log(T/\delta)\right).
\end{aligned}$$

This is exactly the bound claimed in Theorem 4.3.

B Proof of Theorem 4.3

Recall the true backward process is given by

$$\mathrm{d}Y_t = \nabla \log p_{T-t}(Y_t)\mathrm{d}t + \mathrm{d}B_t,$$

and we have defined the estimated backward process by

$$\mathrm{d}\hat{Y}_t = \tilde{s}_t(\hat{Y}_t)\mathrm{d}t + \mathrm{d}B_t,$$

where the drift $\tilde{s}_t$ is defined by

$$\tilde{s}_t(\exp_{\hat{Y}_{t_k}}(v)) := \tilde{s}_k(\exp_{\hat{Y}_{t_k}}(v)) := \eta_\omega(|v|)\mathrm{d}\left[\exp_{\hat{Y}_{t_k}}(v)\right]\left(\hat{s}_t(\hat{Y}_{t_k})\right), \quad t \in [t_k, t_{k+1}),$$

which depends on the process $(\hat{Y}_s)_{t_0\le s\le t}$ before t. Here, we select the scale parameter as $\omega := \frac{c}{\sqrt{d}}$ where the small constant $c > 0$ only depends on r_i and C_{Rm}, and the cut-off function η_ω satisfies

$$\eta_\omega(r) = \begin{cases} 1, & r \le \omega/2, \\ 0, & r \ge \omega, \end{cases} \qquad \eta'_\omega(r) = O(\omega^{-1}), \quad \eta''_\omega(r) = O(\omega^{-2}).$$

The function $\tilde{s}_t$ expand the estimated score function at Y_{t_k} to the local geodesic ball. Note that according to the definition, the vector $\mathrm{d}\left[\exp_{\hat{Y}_{t_k}}(v)\right]\left(\hat{s}_t(\hat{Y}_{t_k})\right)$ is literally a constant vector in the local normal coordinate representation at $\hat{X}_{t_k}$. This can be viewed as a counterpart of Euclidean diffusion model case, where we fix the score function and run the Euler-Maruyama discretization to sample data.

Also, recall that in this proof, our eventual goal is to bound the KL divergence between the distribution of Y_{t_N} and $\hat{Y}_{t_N}$.

B.1 Overall Framework

We adapt the same high-level idea of Xu et al. (2026), which applies the Girsanov's theorem to compute the KL divergence between processes. Let P and $\hat{P}$ be the path measure of $(Y_t)_{t_0\le t\le t_N}$ and $(\hat{Y}_t)_{0\le t\le t_N}$, then, according to the standard Girsanov-based framework (Chen et al., 2023b; De Bortoli et al., 2022),

$$\mathrm{KL}(P\|\hat{P}) = \mathrm{KL}(P_{t_0}\|\hat{P}_{t_0}) + \frac{1}{2}\int_{t_0}^{t_N} \mathbb{E}_{Y\sim P}\left[\|\nabla \log p_{T-t}(Y_t) - \tilde{s}_t(Y_t)\|^2\right] \mathrm{d}t. \tag{11}$$

We now need to compare the two drifts $\nabla \log p_{T-t}(Y_t)$ and $\tilde{s}_t(Y_t)$. We first bridge the two drifts through an intermediate function defined by

$$\tilde{s}^*_t(\exp_{Y_{t_k}}(v)) := \eta_\omega(|v|)\mathrm{d}\left[\exp_{Y_{t_k}}(v)\right](\nabla \log p_{T-t_k}(Y_{t_k}))\,.$$

That is, $\tilde{s}^*_t(\cdot)$ expands the true score function instead of the estimated score function. Clearly,

$$\|\nabla \log p_{T-t}(Y_t) - \tilde{s}_t(Y_t)\|^2 \le 2\underbrace{\|\nabla \log p_{T-t}(Y_t) - \tilde{s}^*_t(Y_t)\|^2}_{\text{Discretization Error}} + 2\underbrace{\|\tilde{s}^*_t(Y_t) - \tilde{s}_t(Y_t)\|^2}_{\text{Score Estimation Error}}.$$

According to Lemma E.5, as long as we take $\omega = \frac{c}{\sqrt{d}}$ for a small global constant depending on C_{Rm} and r_i, the matrix representation of Riemannian metric g inside the local normal coordinate chart with radius ω satisfies $\|g\| \le C$. Therefore, by that $\eta_\omega(\cdot) \le 1$,

$$\|\tilde{s}^*_t(Y_t) - \tilde{s}_t(Y_t)\|^2 \le C\|\hat{s}_{t_k}(Y_{t_k}) - \nabla \log p_{T-t_k}(Y_{t_k})\|^2.$$

Applying Assumption 4.2,

$$\int_{t_0}^{t_N} \mathbb{E}_{Y\sim P}\left[\|\tilde{s}^*_t(Y_t) - \tilde{s}_t(Y_t)\|^2\right] \mathrm{d}t \lesssim \sum_{k=0}^{N-1} h_k \mathbb{E}_{y\sim p_{T-t_k}}\left[\|\hat{s}_{t_k}(y) - \nabla \log p_{T-t_k}(y)\|^2\right] \lesssim \epsilon^2_{\text{score}}.$$

Therefore, in order to prove Theorem 4.3, the only thing left is to control

$$\text{Discretization Error} = 2\int_{t_0}^{t_N} \mathbb{E}\left[\|\nabla \log p_{T-t}(Y_t) - \tilde{s}^*_t(Y_t)\|^2\right] \mathrm{d}t.$$

B.2 Itô's Rule for Discretization Error

In this section, we explicitly compute the formula for discretization error by applying Ito's lemma on $\|\nabla \log p_{T-t}(Y_t) - \tilde{s}^*_t(Y_t)\|^2$ in the interval $t \in [t_k, t_{k+1})$, where we can view Y_{t_k} as a fixed point. To do so, we must compute the time and spacial derivatives of $\|\nabla \log p_{T-t}(Y_t) - \tilde{s}^*_t(Y_t)\|^2$. First, the density function $p_t(\cdot)$ satisfies the heat equation

$$\partial_t p_t(x) = \frac{1}{2}\Delta p_t(x).$$

Hence,

$$\begin{aligned}\partial_t \nabla \log p_t(x) &= \nabla\left(\frac{\partial_t p_t(x)}{p_t(x)}\right) = \frac{1}{2}\nabla\left(\frac{\Delta p_t(x)}{p_t(x)}\right) \\ &\overset{\text{(a)}}{=} \frac{1}{2}\nabla\left(\Delta \log p_t(x) + \|\nabla \log p_t(x)\|^2\right) \\ &= \frac{1}{2}\nabla\Delta \log p_t(x) + \nabla^2 \log p_t(x)\cdot \nabla \log p_t(x),\end{aligned} \tag{12}$$

where step (a) comes from the equality

$$\Delta \log f = \operatorname{div}(\nabla \log f) = \operatorname{div}\left(\frac{\nabla f}{f}\right) = \frac{\Delta f}{f} - \frac{\|\nabla f\|^2}{f^2}.$$

Moreover, $\partial_t \tilde{s}_t^*(\cdot) = 0$ for $t \in [t_k, t_{k+1})$, as the expanded drift is actually independent of t. Therefore, using that $\mathrm{d}Y_t = \nabla \log p_{T-t}(Y_t) + \mathrm{d}B_t$,

$$\begin{aligned}&\mathrm{d}\|\nabla \log p_{T-t}(Y_t) - \tilde{s}_t^*(Y_t)\|^2 \\ &= \partial_t \|\nabla \log p_{T-t}(Y_t) - \tilde{s}_t^*(Y_t)\|^2 \mathrm{d}t + \frac{1}{2}\Delta\|\nabla \log p_{T-t}(Y_t) - \tilde{s}_t^*(Y_t)\|^2 \mathrm{d}t \\ &\quad + \nabla\|\nabla \log p_{T-t}(Y_t) - \tilde{s}_t^*(Y_t)\|^2 \cdot (\nabla \log p_{T-t}(Y_t)\mathrm{d}t + \mathrm{d}B_t) \\ &\overset{\text{(a)}}{=} 2(\nabla \log p_{T-t}(Y_t) - \tilde{s}_t^*(Y_t)) \\ &\quad \cdot \Big[\partial_t \nabla \log p_{T-t}(Y_t)\mathrm{d}t + \nabla^2 \log p_{T-t}(Y_t)\cdot(\nabla \log p_{T-t}(Y_t)\mathrm{d}t + \mathrm{d}B_t) \\ &\qquad - \nabla \tilde{s}_t^*(Y_t)\cdot(\nabla \log p_{T-t}(Y_t) + \mathrm{d}B_t)\Big] \\ &\quad + (\nabla \log p_{T-t}(Y_t) - \tilde{s}_t^*(Y_t))\cdot(\Delta\nabla \log p_{T-t}(Y_t)) - \Delta\tilde{s}_t^*(Y_t))\mathrm{d}t \\ &\quad + \|\nabla^2 \log p_{T-t}(Y_t) - \nabla \tilde{s}_t^*(Y_t)\|_{\text{HS}}^2 \mathrm{d}t \\ &\overset{\text{(b)}}{=} (\nabla \log p_{T-t}(Y_t) - \tilde{s}_t^*(Y_t)) \\ &\quad \cdot \Big[(\Delta\nabla - \nabla\Delta)\log p_{T-t}(Y_t)\mathrm{d}t + \nabla^2 \log p_{T-t}(Y_t)\cdot \mathrm{d}B_t \\ &\qquad - \nabla \tilde{s}_t^*(Y_t)\cdot(\nabla \log p_{T-t}(Y_t)\mathrm{d}t + \mathrm{d}B_t) - \Delta\tilde{s}_t(Y_t)\Big] \\ &\quad + \|\nabla^2 \log p_{T-t}(Y_t) - \nabla \tilde{s}_t^*(Y_t)\|_{\text{HS}}^2 \mathrm{d}t \\ &\overset{\text{(c)}}{=} (\nabla \log p_{T-t}(Y_t) - \tilde{s}_t^*(Y_t)) \\ &\quad \cdot \Big[\mathrm{Ric}^{\#}(\nabla \log p_{T-t}(Y_t), \cdot)\mathrm{d}t + \nabla^2 \log p_{T-t}(Y_t)\cdot \mathrm{d}B_t \\ &\qquad - \nabla \tilde{s}_t^*(Y_t)\cdot(\nabla \log p_{T-t}(Y_t)dt + dB_t) - \Delta\tilde{s}_t(Y_t)\Big] \\ &\quad + \|\nabla^2 \log p_{T-t}(Y_t) - \nabla \tilde{s}_t^*(Y_t)\|_{\text{HS}}^2 \mathrm{d}t.\end{aligned}$$

Here, (a) comes from the equality

$$\Delta\|V\|^2 = 2\langle \Delta V, V\rangle + 2\|\nabla V\|_{\text{HS}}^2,$$

(b) applies Equation 12, and (c) uses the Bochner's identity

$$(\Delta\nabla - \nabla\Delta)f = \mathrm{Ric}^{\#}(\nabla f, \cdot),$$

and $\mathrm{Ric}^{\#}(V, \cdot)$ represents the vector V' satisfying $\langle V', \cdot\rangle = \mathrm{Ric}(V, \cdot)$, i.e., raising one index in the Ricci curvature.

B.3 Controlling Local Part of Discretization Error

Let τ_k be the first time when $\rho(Y_t, Y_{t_k}) \geq r := \omega/2$. Then, by taking expectation and apply the stopping time theorem,

$$\begin{aligned}
&\mathbb{E}\|\nabla \log p_{T-t\wedge\tau_k}(Y_{t\wedge\tau_k}) - \tilde{s}^*_{t\wedge\tau_k}(Y_{t\wedge\tau_k})\|^2 \\
&= \int_{t_k}^t \mathbb{E}\Bigg[1_{u\leq\tau_k}\cdot(\nabla \log p_{T-u}(Y_u) - \tilde{s}^*_u(Y_u)) \\
&\cdot\left(\mathrm{Ric}^{\#}(\nabla \log p_{T-u}(Y_u),\cdot) - \nabla\tilde{s}^*_u(Y_u)\cdot\nabla\log p_{T-u}(Y_u) - \Delta\tilde{s}^*_u(Y_u)\right)\Bigg]\mathrm{d}u \\
&\quad + \int_{t_k}^t \mathbb{E}\left[1_{u\leq\tau_k}\cdot\|\nabla^2\log p_{T-u}(Y_u) - \nabla\tilde{s}^*_t(Y_u)\|^2_{\mathrm{HS}}\right]\mathrm{d}u \\
&\leq \int_{t_k}^t \frac{1}{h_k}\cdot\mathbb{E}\|\nabla\log p_{T-u}(Y_u) - \tilde{s}^*_u(Y_u)\|^2\mathrm{d}u \\
&+ \int_{t_k}^t \frac{h_k}{4}\mathbb{E}\left(1_{u\leq\tau_k}\left\|\mathrm{Ric}^{\#}(\nabla\log p_{T-u}(Y_u),\cdot) - \nabla\tilde{s}^*_u(Y_u)\cdot\nabla\log p_{T-u}(Y_u) - \Delta\tilde{s}^*_u(Y_u)\right\|^2\right)\mathrm{d}u \\
&\quad + \int_{t_k}^t \mathbb{E}\left[1_{u\leq\tau_k}\cdot\|\nabla^2\log p_{T-u}(Y_u) - \nabla\tilde{s}^*_t(Y_u)\|^2_{\mathrm{HS}}\right]\mathrm{d}u,
\end{aligned}$$

where the last inequality directly uses $xy \leq \frac{x^2+y^2}{2}$. Since $\int_{t_k}^t \frac{1}{h_k} \leq 1$, using Grönwall's inequality,

$$\begin{aligned}
&\mathbb{E}\|\nabla \log p_{T-t\wedge\tau_k}(Y_{t\wedge\tau_k}) - \tilde{s}^*_{t\wedge\tau_k}(Y_{t\wedge\tau_k})\|^2 \\
&\leq C\int_{t_k}^t h_k\mathbb{E}\left(1_{u\leq\tau_k}\|\mathrm{Ric}^{\#}(\nabla\log p_{T-u}(Y_u),\cdot) - \nabla\tilde{s}^*_u(Y_u)\cdot\nabla\log p_{T-u}(Y_u) - \Delta\tilde{s}^*_u(Y_u)\|^2\right)\mathrm{d}u \\
&\quad + C\int_{t_k}^t \mathbb{E}\|\nabla^2\log p_{T-u}(Y_u)\|^2_{\mathrm{HS}}\mathrm{d}u + C\int_{t_k}^t \mathbb{E}[\|\nabla\tilde{s}^*_t(Y_u)\|^2_{\mathrm{HS}}1_{u\leq\tau_k}]\mathrm{d}u.
\end{aligned}$$

Now we calculate the term that are related to the locally fixed score $\tilde{s}^*$ within radius $r := \omega/2$. For $x \in \mathcal{M}$ with $\rho(x, Y_{t_k}) \leq r$, consider any local normal coordinate chart at Y_{t_k}, since the coordinate representation of $\tilde{s}^*_t(x)$ is constant inside the local coordinate chart, we have

$$\begin{aligned}
\nabla_\alpha\tilde{s}^*_t(x) &= \Gamma^\beta_{\alpha\gamma}\nabla\log p_{T-t_k}(Y_{t_k})^\gamma\partial_\beta, \\
\Delta\tilde{s}^*_t(x) &= g^{\alpha\beta}(\nabla_\alpha\nabla_\beta - \nabla_{\nabla\alpha(\partial_\beta)})\tilde{s}^*_t(x) \\
&= g^{\alpha\beta}\left[\nabla_\alpha(\Gamma^\xi_{\beta\zeta}\nabla\log p_{T-t_k}(x)^\zeta\partial_\xi) - \Gamma^\xi_{\alpha\beta}\nabla_\xi\tilde{s}^*_t(x)\right] \\
&= g^{\alpha\beta}\left[\partial_\alpha\Gamma^\gamma_{\beta\zeta} + \Gamma^\xi_{\beta\zeta}\Gamma^\gamma_{\alpha\xi} - \Gamma^\xi_{\alpha\beta}\Gamma^\gamma_{\xi\zeta}\right]\nabla\log p_{T-t_k}(Y_{t_k})^\zeta\partial_\gamma.
\end{aligned}$$

Since Lemma E.5 holds for arbitrary choice of local normal coordinates, we can eventually choose a normal coordinate so that $\nabla\log p_{T-t_k}(Y_{t_k})$ has only one nonzero component in the coordinate representation. Therefore, using Lemma E.5,

$$\begin{aligned}
\|\nabla\tilde{s}^*_u(Y_u)\cdot\nabla\log p_{T-u}(Y_u)\| &\leq \|\nabla\log p_{T-t_k}(Y_{t_k})\|\cdot\sqrt{d}\max\Gamma^\beta_{\alpha\gamma}\cdot|\nabla\log p_{T-u}(Y_u)|_1 \\
&\leq C\sqrt{d}\|\nabla\log p_{T-t_k}(Y_{t_k})\|\cdot\|\nabla\log p_{T-u}(Y_u)\|.
\end{aligned}$$

Moreover, Lemma E.5 ensures that $\sum_{\alpha=1}^d\sum_{\beta=1}^d g^{\alpha\beta} = O(d)$ when $\omega \leq \frac{c}{\sqrt{d}}$. So, in the same coordinate setting, for $\rho(x, Y_{t_k}) \leq r$,

$$\begin{aligned}
\|\Delta\tilde{s}^*_t(x)\| &\leq C\sqrt{d}\left[d\cdot\max\partial_\alpha\Gamma^\gamma_{\beta\xi} + d^2(\max\Gamma^\gamma_{\alpha\beta})^2\right]\|\nabla\log p_{T-t_k}(Y_{t_k})\| \\
&\leq C\|\nabla\log p_{T-t_k}(Y_{t_k})\|d^{3/2}.
\end{aligned}$$

In addition, the Ricci curvature is trivially bounded by $d\cdot C_{\mathrm{Rm}}$, so we can bound

$$\|\mathrm{Ric}^{\#}(\nabla\log p_{T-u}(Y_u),\cdot)\| \leq Cd\|\nabla\log p_{T-u}(Y_u)\|.$$

For the Hilbert-Schmidt norm of $\nabla \tilde{s}_t(Y_u)$, apply the local coordinate argument once again, and use the coordinate representation of $\nabla \tilde{s}^*_t(x)$, we obtain that

$$\begin{aligned}\|\nabla \tilde{s}^*_u(x)\|^2_{\text{HS}} &= \|\nabla \log p_{T-t_k}(Y_{t_k})\|^2 \max_{\xi} g^{\alpha\beta} g_{\gamma\eta} \Gamma^{\gamma}_{\alpha\xi} \Gamma^{\eta}_{\beta\xi} \\ &\le C\|\nabla \log p_{T-t_k}(Y_{t_k})\|^2 d^2 \cdot (\max \Gamma)^2 \\ &\le Cd^2 \|\nabla \log p_{T-t_k}(Y_{t_k})\|^2 \rho(x, Y_{t_k})^2.\end{aligned}$$

Combining the results together,

$$\begin{aligned}&\mathbb{E}\|\nabla \log p_{T-t\wedge\tau_k}(Y_{t\wedge\tau_k}) - \tilde{s}^*_{t\wedge\tau_k}(Y_{t\wedge\tau_k})\|^2 \\ \le\ & C\int_{t_k}^{t} h_k \Big(d^2 \mathbb{E}\|\nabla \log p_{T-u}(Y_u)\|^2 \\ &+ d\mathbb{E}\left[\|\nabla \log p_{T-u}(Y_u)\|^2 \cdot \|\nabla \log p_{T-t_k}(Y_{t_k})\|^2\right] + \mathbb{E}\|\nabla \log p_{T-t_k}(Y_{t_k})\|^2 d^3 \Big) \mathrm{d}u \\ &+ C\int_{t_k}^{t} \mathbb{E}\|\nabla^2 \log p_{T-u}(Y_u)\|^2_{\text{HS}} \mathrm{d}u + Cd^2 \int_{t_k}^{t} \mathbb{E}[\|\nabla \log p_{T-t_k}(Y_u)\|^2 \rho(Y_u, Y_{t_k})^2] \mathrm{d}u \\ \overset{\text{(a)}}{\le}\ & C\int_{t_k}^{t} h_k \left(\frac{d^3}{T-u} + \frac{d^4}{T-t_k} + d\mathbb{E}\|\nabla \log p_{T-u}(Y_u)\|^4 + d\mathbb{E}\|\nabla \log p_{T-t_k}(Y_{t_k})\|^4 \right) \mathrm{d}u \\ &+ C\int_{t_k}^{t} \mathbb{E}\|\nabla^2 \log p_{T-u}(Y_u)\|^2_{\text{HS}} \mathrm{d}u + Cd^2 \int_{t_k}^{t} \sqrt{\mathbb{E}\|\nabla \log p_{T-t_k}(Y_{t_k})\|^4} \sqrt{\mathbb{E}\rho(Y_u, Y_{t_k})^4} \mathrm{d}u \\ \overset{\text{(b)}}{\le}\ & C\int_{t_k}^{t} h_k \left(\frac{d^3}{T-u} + \frac{d^4}{T-t_k} + \frac{d^3}{(T-u)^2} + \frac{d^3}{(T-t_k)^2} \right) \mathrm{d}u \\ &+ C\int_{t_k}^{t} \mathbb{E}\|\nabla^2 \log p_{T-u}(Y_u)\|^2_{\text{HS}} \mathrm{d}u + Cd^2 \int_{t_k}^{t} \frac{d}{T-t_k} \cdot du \cdot \mathrm{d}u,\end{aligned}$$

where we apply Lemma E.2 for to obtain inequality (a), and apply Lemma E.3 to obtain inequality (b). Note that during step (b), we used the fact that (Y_t) is exactly the time reversal of the forward process (X_t), and the proof of Lemma E.3 shows that the radial process $r(X_t)$ of Riemmannian Brownian motion (under nonnegative Ricci curvature) satisfies $\mathbb{E}[r(X_t)^4] \le O(d^2t^2)$.

Since we want to control the total discretization error, we must take integration over t. Thus,

$$\begin{aligned}\text{Discretization Error} &= C\int_{t_0}^{t_N} \mathbb{E}\left[\|\nabla \log p_{T-t}(Y_t) - \tilde{s}^*_t(Y_t)\|^2\right] \mathrm{d}t \\ &\le C\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}\|\nabla \log p_{T-t\wedge\tau_k}(Y_{t\wedge\tau_k}) - \tilde{s}^*_{t\wedge\tau_k}(Y_{t\wedge\tau_k})\|^2 \mathrm{d}t \\ &\quad + C\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}\left[\|\nabla \log p_{T-t}(Y_t) - \tilde{s}^*_t(Y_t)\|^2 \cdot 1_{\tau_k<t}\right] \mathrm{d}t \\ &\le C\sum_{k=0}^{N-1} h_k^3 \left(\frac{d^4}{T-t_{k+1}} + \frac{d^3}{(T-t_{k+1})^2} \right) \\ &\quad \underbrace{+ C\sum_{k=0}^{N-1} h_k \int_{t_k}^{t_{k+1}} \mathbb{E}\|\nabla^2 \log p_{T-t}(Y_t)\|^2_{\text{HS}} \mathrm{d}t}_{\text{Quadratic Variation Term}} \\ &\quad \underbrace{+ C\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \sqrt{\mathbb{E}(\|\nabla \log p_{T-t}(Y_t)\| + \|\tilde{s}^*_t(Y_t)\|)^4} \cdot \sqrt{P(\tau_k < t)} \mathrm{d}t}_{\text{Tail Event Term}}.\end{aligned}$$

Here, the first term comes from the analyses in this section, which has been already calculated. The second term comes from the quadratic variation of the reverse process, and the third term quantifies

the contribution of tail events, where the reverse process goes outside the local coordinate chart. In the following sections, we will eventually control the second and third term.

B.4 Handle the Probability of Tail Events

In this section, we provide upper bounds for the tail events term in the discretization error. First, according to Lemma E.5, within the local coordinate chart at Y_{t_k} with radius $\omega = \frac{c}{\sqrt{d}}$, the metric tensor satisfies $\|g(x)\| \le 2$. Therefore, by Lemma E.3,

$$\begin{aligned}
&\sqrt{\mathbb{E}(\|\nabla \log p_{T-t}(Y_t)\| + \|\tilde{s}_t^*(Y_t)\|)^4} \\
&\le C\sqrt{\mathbb{E}\|\nabla \log p_{T-t}(Y_t)\|^4} + C\sqrt{\mathbb{E}\|\nabla \log p_{T-t_k}(Y_t)\|^4} \\
&\le \frac{Cd}{T - t_{k+1}}.
\end{aligned}$$

Therefore, we only need to control the probability that the process Y_t goes outside the local coordinate chart. Define the distance function $r(x) := \rho(Y_{t_k}, x)$, which is smooth inside the local coordinate chart. Moreover, by Laplacian comparison theorem (see, e.g., Theorem 3.4.2 of Hsu 2002), when the manifold $\mathcal{M}$ has nonnegative Ricci curvature, $\Delta r^2(x) \le 2d$. Then, using Itô's rule, for any integer $m \ge 2$,

$$\begin{aligned}
r(Y_{t\wedge\tau_k})^{2m} &= \int_{t_k}^{t\wedge\tau_k} \left\{ \nabla \left[r(Y_u)^{2m} \right] \cdot (\nabla \log p_{T-u}(Y_u)\mathrm{d}u + \mathrm{d}B_u) + \frac{1}{2}\Delta \left[r(Y_u)^{2m} \right] \mathrm{d}u \right\} \\
&\le \int_{t_k}^{t\wedge\tau_k} \left(2m \cdot r(Y_u)^{2m-1} \|\nabla \log p_{T-u}(Y_u)\| + \frac{1}{2}\mathrm{div}(m r(Y_u)^{2m-2} \nabla r(Y_u)^2) \right) \mathrm{d}u \\
&\quad + 2m \int_{t_k}^{t\wedge\tau_k} r(Y_u)^{2m-1} \cdot \mathrm{d}\beta_u \\
&\le \int_{t_k}^{t\wedge\tau_k} \left(2m \cdot r(Y_u)^{2m-1} \|\nabla \log p_{T-u}(Y_u)\| + (md + (2m(m-1)) r(Y_u)^{2m-2} \right) \mathrm{d}u \\
&\quad + 2m \int_{t_k}^{t\wedge\tau_k} r(Y_u)^{2m-1} \cdot \mathrm{d}\beta_u,
\end{aligned}$$

where (β_t) is a one-dimensional Brownian motion. Taking expectation,

$$\begin{aligned}
\mathbb{E}\left[r(Y_{t\wedge\tau_k})^{2m} \right] &\le \int_{t_k}^{t} \mathbb{E}\left[r(Y_{u\wedge\tau_k})^{2m-1} \cdot C_m \|\nabla \log p_{T-u}(Y_u)\| \right] \mathrm{d}u \\
&\quad + \int_{t_k}^{t} \mathbb{E}\left[r(Y_{u\wedge\tau_k})^{2m-2} \cdot C_m d \| \right] \mathrm{d}u \\
&\overset{\text{(a)}}{\le} \int_{t_k}^{t} \left(\frac{\mathbb{E}[(r(Y_{u\wedge\tau_k})^{2m-1} / h_k^{(2m-1)/2m})^{2m/(2m-1)}]}{2m/(2m-1)} \right. \\
&\qquad \left. + \frac{\mathbb{E}[(C_m \|\nabla \log p_{T-u}(Y_u)\| h_k^{(2m-1)/2m})^{2m}]}{2m} \right) \mathrm{d}u \\
&\quad + \int_{t_k}^{t} \left(\frac{\mathbb{E}[(r(Y_{u\wedge\tau_k})^{2m-2} / h_k^{(m-1)/m})^{m/(m-1)}]}{m/(m-1)} \right. \\
&\qquad \left. + \frac{\mathbb{E}[(C_m d h_k^{(m-1)/m})^m]}{m} \right) \mathrm{d}u \\
&\le \int_{t_k}^{t} \frac{2}{h_k} \mathbb{E}[r(Y_{u\wedge\tau_k})^{2m}] \mathrm{d}u \\
&\quad + C_m h_k^{2m-1} \int_{t_k}^{t} \mathbb{E}\|\nabla \log p_{T-u}(Y_u)\|^{2m} \mathrm{d}u + C_m h_k^m d^m,
\end{aligned}$$

where step (a) uses Young's inequality $xy \le \frac{x^p}{p} + \frac{x^q}{q}$ for (p, q) satisfying $p > 1, q > 1$ and $p^{-1} + q^{-1} = 1$. Since $\int_{t_k}^{t} \frac{2}{h_k} \mathrm{d}u \le 2$, by Grönwall's inequality,

$$\mathbb{E}\left[r(Y_{t\wedge\tau_k})^{2m}\right] \le C_m h_k^m d^m + C_m h_k^{2m-1} \int_{t_k}^{t} \mathbb{E}\|\nabla \log p_{T-u}(Y_u)\|^{2m} \mathrm{d}u.$$

Take $m = 6$, then

$$\mathbb{E}\left[r(Y_{t\wedge\tau_k})^{12}\right] \le C h_k^6 d^6 + C h_k^{12} d^6/(T - t_{k+1})^6 \le C h_k^6 d^6,$$

if we assume that $h_k = O((T - t_{k+1})^{-1})$. Hence, by Markov's inequality,

$$\begin{aligned} P(\tau_k < t) &= P(r(Y_{t\wedge\tau_k}) \ge \omega/2) \\ &\le \frac{\mathbb{E}[r(Y_{t\wedge\tau_k})^{12}]}{(\omega/2)^{12}} \le C h_k^6 d^{12}. \end{aligned}$$

Therefore, we conclude that

$$\text{Tail Event Term} \le C \sum_{k=0}^{N-1} h_k \cdot \frac{d}{T - t_{k+1}} \cdot h_k^3 d^6 \le C \sum_{k=0}^{N-1} \frac{h_k^4 d^7}{T - t_{k+1}}.$$

B.5 Sharp Analysis of Quadratic Variation Term

In this section, we provide a sharp analysis for the following part of discretization error:

$$\text{Quadratic Variation Term} = C \sum_{k=0}^{N-1} h_k \int_{t_k}^{t_{k+1}} \mathbb{E}\|\nabla^2 \log p_{T-t}(Y_t)\|_{\text{HS}}^2 \mathrm{d}t.$$

As we will show in this section, this is the dominant term of discretization error, and it has a linear growth in d.

First, note that $\|\nabla \log p_{T-t}(x)\|^2$ is a smooth function in $x \in \mathcal{M}$. Hence, apply Itô's lemma again on $\mathrm{d}Y_t = \nabla \log p_{T-t}(Y_t)\mathrm{d}t + \mathrm{d}B_t$,

$$\begin{aligned} &\mathrm{d}\|\nabla \log p_{T-t}(Y_t)\|^2 \\ &= \partial_t \|\nabla \log p_{T-t}(Y_t)\|^2 \mathrm{d}t + \nabla \|\nabla \log p_{T-t}(Y_t)\|^2 \cdot \nabla \log p_{T-t}(Y_t) \mathrm{d}t \\ &\quad + \frac{1}{2}\Delta \|\nabla \log p_{T-t}(Y_t)\|^2] \mathrm{d}t + \nabla \|\nabla \log p_{T-t}(Y_t)\|^2 \cdot \mathrm{d}B_t \\ &= 2\nabla \log p_{T-t}(Y_t) \cdot \partial_t \nabla \log p_{T-t}(Y_t) \mathrm{d}t \\ &\quad + 2\nabla \log p_{T-t}(Y_t) \cdot (\nabla^2 \log p_{T-t}(Y_t) \cdot \nabla \log p_{T-t}(Y_t)) \mathrm{d}t \\ &\quad + \nabla \log p_{T-t}(Y_t) \cdot \Delta \nabla \log p_{T-t}(Y_t) \mathrm{d}t + \|\nabla \log p_{T-t}(Y_t)\|_{\text{HS}}^2 \mathrm{d}t \\ &\quad + \nabla \|\nabla \log p_{T-t}(Y_t)\|^2 \cdot \mathrm{d}B_t \\ &= \nabla \log p_{T-t}(Y_t) \cdot \Big[-2\nabla^2 \log p_{T-t}(Y_t) \cdot \nabla \log p_{T-t}(Y_t) - \nabla \Delta \log p_{T-t}(Y_t) \\ &\quad + 2\nabla^2 \log p_{T-t}(Y_t) \cdot \nabla \log p_{T-t}(Y_t) + \Delta \nabla \log p_{T-t}(Y_t) \Big] \mathrm{d}t \\ &\quad + \|\nabla \log p_{T-t}(Y_t)\|_{\text{HS}}^2 \mathrm{d}t + \nabla \|\nabla \log p_{T-t}(Y_t)\|^2 \cdot \mathrm{d}B_t \\ &= \mathrm{Ric}(\nabla \log p_{T-t}(Y_t), \nabla \log p_{T-t}(Y_t)) \mathrm{d}t + \|\nabla \log p_{T-t}(Y_t)\|_{\text{HS}}^2 \mathrm{d}t \\ &\quad + \nabla \|\nabla \log p_{T-t}(Y_t)\|^2 \cdot \mathrm{d}B_t. \end{aligned}$$

Take expectation,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\|\nabla \log p_{T-t}(Y_t)\|^2 &= \mathbb{E}\mathrm{Ric}(\nabla \log p_{T-t}(Y_t), \nabla \log p_{T-t}(Y_t)) + \mathbb{E}\|\nabla \log p_{T-t}(Y_t)\|_{\text{HS}}^2 \\ &\ge \mathbb{E}\|\nabla \log p_{T-t}(Y_t)\|_{\text{HS}}^2, \end{aligned}$$

where the last inequality uses the assumption that $\mathcal{M}$ has nonnegative Ricci curvature. By product rule, we can immediately obtain that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left((T-t)\mathbb{E}\|\nabla \log p_{T-t}(Y_t)\|^2\right) &= (T-t)\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\|\nabla \log p_{T-t}(Y_t)\|^2 - \mathbb{E}\|\nabla \log p_{T-t}(Y_t)\|^2 \\ &\geq (T-t)\mathbb{E}\|\nabla \log p_{T-t}(Y_t)\|_{\mathrm{HS}}^2 - \mathbb{E}\|\nabla \log p_{T-t}(Y_t)\|^2.\end{aligned}$$

Now, assume that the step sizes are chosen so that $h_k = O(h \cdot (T - t_{k+1}))$, then, let $\delta = T - t_N$,

$$\begin{aligned}&\text{Quadratic Variation Term} \\ &= C\sum_{k=0}^{N-1} h_k \int_{t_k}^{t_{k+1}} \mathbb{E}\|\nabla^2 \log p_{T-t}(Y_t)\|_{\mathrm{HS}}^2 \mathrm{d}t. \\ &\leq C\sum_{k=0}^{N-1} h \int_{t_k}^{t_{k+1}} (T-t)\mathbb{E}\|\nabla^2 \log p_{T-t}(Y_t)\|_{\mathrm{HS}}^2 \mathrm{d}t \\ &\leq Ch\int_0^{t_N} \left[\frac{\mathrm{d}}{\mathrm{d}t}\left((T-t)\mathbb{E}\|\nabla \log p_{T-t}(Y_t)\|^2\right) + \mathbb{E}\|\nabla \log p_{T-t}(Y_t)\|^2\right] \mathrm{d}t \\ &\leq Ch\left[\delta\mathbb{E}\|\nabla \log p_\delta(Y_{T-\delta})\|^2 - T\mathbb{E}\|\nabla \log p_T(Y_0)\|^2 + \int_\delta^T \mathbb{E}\|\nabla \log p_t(X_t)\|^2 \mathrm{d}t\right] \\ &\leq Ch\left[\delta\mathbb{E}\|\nabla \log p_\delta(X_\delta)\|^2 + \int_\delta^T \frac{d}{t}\mathrm{d}t\right] \\ &\leq Ch\left[d + d\log(T/\delta)\right] \leq Chd\log(T/\delta),\end{aligned}$$

where we have used Lemma E.2 to control the second moment of the score function.

B.6 Combining the Results

We define the step size by $h_k = \min\{(T-\delta)-t_k, h\cdot T - t_k\}$, where $h > 0$ is a parameter controlling the scale of step sizes. Then, as long as $h \leq \frac{1}{2}$, the step size satisfies $h_k \leq 2(T - t_{k+1})h$. Moreover, the total number of steps is given by

$$N = \lceil \log_{1-h}(\delta/T)\rceil = \left\lceil \frac{\log(T/\delta)}{-\log(1-h)}\right\rceil = \Theta\left(h^{-1}\log(T/\delta)\right).$$

Now, with this schedule of step sizes,

$$\begin{aligned}\text{Discretization Error} &\leq Ch^2\sum_{k=0}^{N-1} h_k\left(d^4(T-t_{k+1}) + d^3\right) \\ &\quad + Ch^3\sum_{k=0}^{N-1} h_k d^7 (T-t_{k+1})^2 \\ &\quad + Chd\log(T/\delta) \\ &\leq Ch^2\int_\delta^T d^4 t\mathrm{d}t + Ch^3\int_\delta^T d^7 t^2 \mathrm{d}t + Chd\log(T/\delta) \\ &\leq Chd\log(T/\delta) + Ch^2 d^4 T^2 + Ch^3 d^7 T^3.\end{aligned}$$

This proves the error bound claimed in the theorem.

In addition, in order to achieve ϵ^2 discretization error, it suffices to take

$$h^{-1} = \Theta\left(\frac{d\log(T/\delta)}{\epsilon^2} + \frac{d^2 T}{\epsilon} + \frac{d^{7/3}T}{\epsilon^{2/3}}\right),$$

where the corresponding number of steps is

$$N = \Theta\left(\left(\frac{d\log(T/\delta)}{\epsilon^2} + \frac{d^2 T}{\epsilon} + \frac{d^{7/3}T}{\epsilon^{2/3}}\right)\log(T/\delta)\right).$$

When the target ϵ^2 is sufficiently small, the formula can be simplified to

$$N = \Theta\left(\frac{d\log^2(T/\delta)}{\epsilon^2}\right).$$

C Introduction to the Error Bound without Nonnegative Ricc Curvature Assumption

In this section, we briefly introduce how to obtain the discretization error bound when the last assumption in Assumption 4.1 is replaced by a bounded Ricci curvature assumption, where the manifold $\mathcal{M}$ has Ricci curvature bounded below by $-C_R$ for some constant $C_R > 0$. A worst-case estimate gives $C_R = C_{\text{Rm}}d$, but we use C_R here for generality.

Under this assumption, all the error partition in Appendix B holds, and the only difference is that the moment of true score function can be larger (see e.g., Lemma E.4). For simplicity, we only focus on the dominating term

$$\text{Quadratic Variation Term} = C\sum_{k=0}^{N-1} h_k \int_{t_k}^{t_{k+1}} \mathbb{E}\|\nabla^2 \log p_{T-t}(Y_t)\|_{\text{HS}}^2 \mathrm{d}t,$$

since the summation of other discretization error term is of order $o(h)$ for step size scale h.

To analyze the Hilbert-Schmidt term, we further apply Itô's formula on $\|\nabla \log p_{T-t}(Y_t)\|^2$ as in Appendix B.5, but with the new assumption,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\|\nabla \log p_{T-t}(Y_t)\|^2 &= \mathbb{E}\text{Ric}(\nabla \log p_{T-t}(Y_t), \nabla \log p_{T-t}(Y_t)) + \mathbb{E}\|\nabla \log p_{T-t}(Y_t)\|_{\text{HS}}^2 \\ &\geq \mathbb{E}\|\nabla \log p_{T-t}(Y_t)\|_{\text{HS}}^2 - C_R\mathbb{E}\|\nabla \log p_{T-t}(Y_t)\|^2,\end{aligned}$$

Moreover,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left((T-t)\mathbb{E}\|\nabla \log p_{T-t}(Y_t)\|^2\right) &\geq (T-t)\mathbb{E}\|\nabla \log p_{T-t}(Y_t)\|_{\text{HS}}^2 \\ &\quad - (1 + C_R(T-t))\mathbb{E}\|\nabla \log p_{T-t}(Y_t)\|^2,\end{aligned}$$

Now, instead, we define the step sizes so that $h_k := \min\{T - t_k, 1\} \cdot h$, then, using Lemma E.4,

$$\begin{aligned}&\text{Quadratic Variation Term} \\ &= C\sum_{k=0}^{N-1} h_k \int_{t_k}^{t_{k+1}} \mathbb{E}\|\nabla^2 \log p_{T-t}(Y_t)\|_{\text{HS}}^2 \mathrm{d}t. \\ &\leq Ch\left(\int_0^{T-\delta} \min\{T-t,1\}\mathbb{E}\|\nabla^2 \log p_{T-t}(Y_t)\|_{\text{HS}}^2 \mathrm{d}t\right) \\ &\leq Ch\Bigg(-\int_\delta^T \frac{\mathrm{d}}{\mathrm{d}t}\left(\min\{t,1\}\mathbb{E}\|\nabla \log p_t(X_t)\|^2\right)\mathrm{d}t \\ &\quad + C_R\int_\delta^T \min\{t,1\}\mathbb{E}\|\nabla \log p_t(X_t)\|^2 \mathrm{d}t + \int_\delta^1 \mathbb{E}\|\nabla \log p_t(X_t)\|^2 \mathrm{d}t\Bigg) \\ &\leq Ch(d + \sqrt{dC_R}D)\left(\delta\left(\frac{1}{\delta} + C_R\right) + C_R\int_\delta^T (1 + C_R)\,\mathrm{d}t + \int_\delta^1 \left(\frac{1}{t} + C_R\right)\mathrm{d}t\right) \\ &\leq Ch(d + \sqrt{dC_R}D)\left(TC_R + \log(1/\delta)\right).\end{aligned}$$

Note that when we use this step size schedule, the number of step sizes N satisfies $N = O\left(h^{-1}(T + \log(1/\delta))\right)$. Hence, the number of steps required to achieve ϵ^2-KL error satisfies

$$N = O\left(\frac{(d + \sqrt{dC_R}D)(T^2C_R + TC_R\log(1/\delta) + \log^2(1/\delta))}{\epsilon^2}\right) = \tilde{O}\left(\frac{C_RT^2(d + \sqrt{dC_R}D)}{\epsilon^2}\right).$$

In addition, when we use the worst-case estimate $C_R = C_{\mathrm{Rm}}d$, this leads to a bound given by

$$N = \tilde{O}\left(\frac{T^2 d^2}{\epsilon^2}\right).$$

Although it is still significantly better than the $\tilde{O}\left(\frac{T^2 d^6}{\epsilon^2}\right)$ bound provided by Xu et al. (2026), it is worse than the $O\left(\frac{d\log(T/\delta)}{\epsilon^2}\right)$ bound we can obtain under the assumption $C_R = 0$. This further highlights that Riemannian diffusion models do perform better on Riemannian manifolds with nonnegative Ricci curvature.

D Convergence Rate of Geodesic Random Walk

In the previous section, we have controlled the KL divergence between an estimated process $(\hat{Y}_t)_{0\le t\le T-\delta}$ and the true reverse process $(Y_t)_{0\le t\le T-\delta}$. The estimated process is defined by

$$\mathrm{d}\hat{Y}_t = \tilde{s}_t(\hat{Y}_t)\mathrm{d}t + \mathrm{d}B_t,$$

where the drift function $\tilde{s}_k(x) := \tilde{s}_t(x)$ is fixed within each interval $t \in [t_k, t_{k+1})$ given the state $\hat{Y}_{t_k}$ at time t_k.

Now, we use Geodesic Random Walk to approximate the estimated process. Specifically, for each interval $[t_k, t_{k+1})$, we partition the interval into n_{bm} small steps $(t_{k,0} = t_k, t_{k,1}, t_{k,2}, \dots, t_{k,n_k} = t_{k+1})$. Given $\tilde{Y}_{t_k}$ at time t_k, we sample the state at time t_{k+1} by the following iteration:

$$\tilde{Y}_{t_{k,i+1}} = \exp_{\tilde{Y}_{t_{k,i}}}\left((t_{k,i+1} - t_{k,i})\tilde{s}_k(\tilde{Y}_{t_{k,i}}) + \sqrt{t_{k,i+1} - t_{k,i}}w_{k,i}\right), \quad i = 0, \dots, n_k - 1,$$

where $w_{k,i}$ is an independent Gaussian random vector in the tangent space $T_{\tilde{Y}_{t_{k,i}}}\mathcal{M}$.

Through the Geodesic Random walk, we define a sampling process $\left(\tilde{Y}_{t'_k}\right)_{k=0}^{N'}$ where we use t' to represent the discretization sequence given by $(t_{k,i})$ for notational simplicity. Define $\tilde{P}_i(y_{i+1}|(y_j)_{j\le i})$ as the transition kernel of process $\tilde{Y}$, that is, the conditional density of $\tilde{Y}_{t'_{i+1}}$ given $(\tilde{Y}_{t'_j})_{j\le i}$. Similarly, define $\hat{P}_i(y_{i+1}|(y_j)_{j\le i})$ as the corresponding transition kernel of $(\hat{Y}_{t'_i})_{i=0}^{N'}$. Then,

$$\begin{aligned}\mathrm{KL}\left(\mathrm{Law}(\tilde{Y}_{T-\delta})\middle\|\mathrm{Law}(\hat{Y}_{T-\delta})\right) &\le \mathrm{KL}\left(\mathrm{Law}((\tilde{Y}_{t'_i})_{i=0}^{N'})\middle\|\mathrm{Law}((\hat{Y}_{t'_i})_{i=0}^{N'})\right)\\ &= \sum_{i=0}^{N'}\mathbb{E}\left[\mathrm{KL}(\tilde{P}_i(\cdot|(\tilde{Y}_{t'_j})_{j\le i})\|\hat{P}_i(\cdot|(\tilde{Y}_{t'_j})_{j\le i}))\right].\end{aligned}$$

This provides is general idea of controlling the Brownian-motion simulation error, and the details will be introduced below. In the following sections, we will apply heat kernel theories from Chapter 2 of (Berline et al., 2004) to write the explicit formula for the transition kernel $\hat{P}$. Though many high-level ideas comes from the analysis in (Xu et al., 2026), their work has only argued that there is a polynomial guarantee on the convergence rate, and did not provide the explicit calculation on the degree of d in the polynomial factor.

D.1 Exact Formulation of Transition Kernel.

Consider the interval $[t_k, t_{k+1})$, and we have fixed the drift function $\tilde{s}_k(\cdot)$ in this interval. With the drift, the transition density $\hat{P}(t, x, y)$ satisfies the heat function

$$(\partial_t + H)\hat{P}(t, x, y) = 0,$$

where the generalized Laplacian H is defined as

$$H := -\frac{1}{2}\Delta + \langle\tilde{s}_k(x), \nabla\rangle + \mathrm{div}(\tilde{s}_k(x)).$$

D.1.1 Formal Solution of the Heat Equation.

Assume that the initial point y is fixed. Consider the local normal coordinate chart at y. For each point x inside the chart, define

$$j(x) := \sqrt{\det(g(x))}, \qquad g \text{ is the Riemannian metric matrix.}$$

That is, $j(x)$ represents the volume element around x. Inside the normal coordinate chart, we define the Euclidean normal density as

$$q_t(x,y) := \frac{1}{(2\pi t)^{d/2}} \exp\left(-\frac{\rho(x,y)^2}{2t}\right).$$

Moreover, we also define

$$\hat{q}_t(x,y) := q_t(x,y) \cdot j(x)^{-1/2}.$$

With these definitions in mind, we now introduce the formal solution of the heat equation. Let $x_s = \exp_y(s \cdot \log_y x)$ be the geodesic curve from y to x. Moreover, define the operator B as

$$B := j^{1/2} \circ H \circ j^{-1/2},$$

where the scalar functions are viewed as multiplication operator, and $\circ$ means composition. Then we have the following theorem.

Theorem D.1 (Formal solution, Theorem 2.26 in Berline et al. 2004)**.** *There exists a unique formal solution $k_t(x,y)$ of the heat equation*

$$(\partial_t + H)k_t(x,y) = 0$$

of the form

$$k_t(x,y) = \hat{q}_t(x,y) \sum_{i=0}^{\infty} t^i u_i(x,y),$$

such that $u_0(y,y) = 1$. In addition, the explicit formula for u_i is given by

$$u_0(x,y) = \tau(x,y) := \exp\left(\int_0^1 \langle \dot{x}_s, \tilde{s}_k(x_s)\rangle \mathrm{d}s\right),$$

$$u_i(x,y) = -\tau(x,y) \int_0^1 \tau(x_s,y)^{-1} s^{i-1} B u_{i-1}(x_s,y)\mathrm{d}s, \quad i \geq 1.$$

Theorem D.1 provide explicit formula for the formal solution. In fact, the definition of formal solution and Proposition 2.24 in (Berline et al., 2004) indicates that, if we define

$$k_t^{(n)}(x,y) := \hat{q}_t(x,y) \sum_{i=0}^{n} t^i u_i(x,y)$$

as the clipped formal solution, this solution satisfies

$$(\partial_t + H)k_t^{(n)}(x,y) = t^n \hat{q}_t(x,y)(Bu_n)(x,y).$$

D.1.2 Parametrix Estimate

According to Theorem D.1, we now focus on $k_t^{(1)}(x,y)$, which can be viewed as an approximation of the true heat kernel $\hat{P}$ of generalized Laplacian H. However, $k_t^{(1)}(x,y)$ is only defined locally. To extend $k_t^{(1)}(x,y)$ onto the whole manifold $\mathcal{M}$ and keep the smoothness, we further define

$$\tilde{k}_t^{(1)}(x,y) = \eta_\iota(\rho(x,y))k_t^{(1)}(x,y),$$

where η_ι is the cut-off function with parameter $\iota > 0$ defined later. This definition refers to (Berline et al., 2004), which is also used in (Xu et al., 2026).

By summarizing the conclusions in Theorem 2.19, Theorem 2.23 and Theorem 2.29 in (Berline et al., 2004), we obtain the following exact representation of heat kernel.

Theorem D.2 (Parametrix)**.** *Define the remainder*

$$r_t(x,y) := (\partial_t + H)\tilde{k}_t^{(1)}(x,y).$$

Let $\Delta_n := \{(t_1,\dots,t_n) : 0 \le t_1 \le t_2 \le \cdots \le t_n \le 1\}$ *be the standard simplex. Define*

$$Q^{(n)}(t,x,y) := \int_{t\Delta_n}\int_{\mathcal{M}^n} \tilde{k}^{(1)}_{t-t_n}(x,x_n) r_{t_n-t_{n-1}}(x_n,x_{n-1}) r_{t_{n-1}-t_{n-2}}(x_{n-1},x_{n-2}) \\ \cdots r_{t_2-t_1}(x_2,x_1) r_{t_1}(x_1,y)\mathrm{d}x_n\cdots\mathrm{d}x_1\mathrm{d}t_n\cdots\mathrm{d}t_1,$$

and $Q^{(0)}(t,x,y) := \tilde{k}_t^{(1)}(x,y)$*. Then,*

$$\hat{P}(t,x,y) = \sum_{n=0}^{\infty}(-1)^n Q^{(n)}(t,x,y).$$

By Theorem D.2,

$$\hat{P}(t,x,y) = \hat{q}_t(x,y)\tau(x,y) + t\hat{q}_t(x,y)u_1(x,y) + \sum_{n=1}^{\infty}(-1)^n Q^{(n)}(t,x,y).$$

Since the sampling process actually sample a Gaussian vector in the tangent space, intuitively, the transition kernel $\tilde{P}$ of the sampling process is close to $\hat{q}_t(x,y)\tau(x,y)$. Therefore, to control the difference between $\hat{P}$ and $\tilde{P}$, we need to provide bounds on

- $t\hat{q}_t(x,y)u_1(x,y)$, which will be shown as of order $O(t)$, and
- $\sum_{n=1}^{\infty}(-1)^n Q^{(n)}(t,x,y)$, which will be shown as of order $o(t)$.

D.2 Bounds on u_1 and its Derivatives

In this section we provide bounds on $|u_1(x,y)|$ to control the first-order error term. In addition, we want to provide bounds on the derivatives to prepare for the later calculations.

We first expand the operator B as follows:

$$\begin{aligned} B &= j^{1/2}\circ H\circ j^{-1/2} \\ &= -\frac{1}{2}\Delta + \langle \tilde{s}_k(x) - j^{1/2}\nabla j^{-1/2}, \nabla\rangle \\ &\quad + \mathrm{div}(\tilde{s}_k(x)) + \frac{1}{2}j^{1/2}\Delta j^{-1/2} + \langle V(x), j^{1/2}\nabla j^{-1/2}\rangle. \end{aligned}$$

Moreover, in local coordinates, the Laplace-Beltrami operator can be written as

$$\Delta = g^{\alpha\beta}\left[\partial_{\alpha\beta} - \Gamma^{\gamma}_{\alpha\beta}\partial_\gamma\right].$$

Therefore,

$$\begin{aligned} u_1(x,y) &= -\tau(x,y)\int_0^1 \tau(x_s,y)^{-1} B_{x_s}(\tau(x_s,y))\mathrm{d}s \\ \text{(zeroth-order part)} \quad &= -\tau(x,y)\int_0^1 \left(\mathrm{div}(\tilde{s}_k(x_s)) + \frac{1}{2}j^{1/2}\Delta j^{-1/2} + \langle \tilde{s}_k(x_s), j^{1/2}\nabla j^{-1/2}\rangle\right)\mathrm{d}s \\ \text{(first-order part)} \quad &\quad - \tau(x,y)\int_0^1 \tau(x_s,y)^{-1}\cdot \mathrm{d}s \\ &\quad \cdot \langle \tilde{s}_k(x_s) - j^{1/2}\nabla j^{-1/2} + g^{\alpha\beta}\Gamma^{\gamma}_{\alpha\beta}\partial_\gamma, \nabla_{x_s}\tau(x_s,y)\rangle \\ \text{(second-order part)} \quad &\quad - \tau(x,y)\int_0^1 \tau(x_s,y)^{-1}\cdot \mathrm{d}s\cdot g^{\alpha\beta}\partial_{\alpha\beta}\tau(x_s,y)). \end{aligned}$$

We first consider the zeroth-order part. By applying Lemma E.5 to the local coordinate at $\hat{Y}_{t_k}$,

$$|\mathrm{div}(\tilde{s}_k(x_s))| = |\partial_\alpha \tilde{s}_k^\alpha + \frac{1}{2} g^{\alpha\beta}\partial_\gamma g_{\alpha\beta}\tilde{s}_k^\gamma| \le C\|\hat{s}_{t_k}(\hat{Y}_{t_k})\|(\omega^{-1}1_{\omega/2\le\rho(x_s,\hat{Y}_{t_k})\le\omega} + d),$$

Moreover, by applying Lemma E.5 to the local coordinate chart at y,

$$\begin{aligned}\partial_\alpha j &= (\partial_\alpha g_{\beta\gamma})g^{\beta\gamma} j = j\cdot O(d\rho(x,y)),\\ \partial_{\alpha\beta} j &= (\partial_\alpha g_{\gamma\xi})g^{\gamma\xi}(\partial_\beta g_{\mu\zeta})g^{\mu\zeta} j + (\partial_{\alpha\beta} g_{\gamma\xi})g^{\gamma\xi} j + (\partial_\alpha g_{\gamma\xi})(\partial_\beta g^{\gamma\xi}) j\\ &= j\cdot O((d\rho(x,y))^2 + d + d^2\rho(x,y)^2) = j\cdot O(d).\end{aligned}$$

Thus, as long as $\rho(x,y)\le O(d^{-1/2})$, the zeroth-order part does not exceed

$$\tau(x,y)\cdot O(\|\hat{s}_{t_k}(\hat{Y}_{t_k})\|d + d^2).$$

We now turn to the first-order part. Let $\log_y x = x^\alpha\partial_\alpha$ be the local normal coordinate representation of x. Then,

$$\begin{aligned}\partial_\alpha\tau(x,y) &= \tau(x,y)\partial_\alpha\int_0^1 \langle \dot{x}_s, \tilde{s}_k(x_s)\rangle \mathrm{d}s\\ &= \tau(x,y)\int_0^1 \left[g_{\alpha\beta}\cdot\tilde{s}_k(x_s)^\beta + s x^\xi\left(g_{\xi\beta}\cdot\partial_\alpha\tilde{s}_k(x_s)^\beta + \partial_\alpha g_{\xi\beta}\cdot\tilde{s}_k(x_s)^\beta\right)\right]\mathrm{d}s,\end{aligned}$$

So, as long as $\rho(x,y)\le\iota = O(d^{-1/2})$, apply Lemma E.5 on g and sum over the elements,

$$\begin{aligned}\|\nabla\tau(x,y)\| &\le \tau(x,y)\cdot O\left(\|\hat{s}_{t_k}(\hat{Y}_{t_k})\|(1+\rho(x,y)(\omega^{-1}+d\omega+\rho(x,y)))\right)\\ &\le \tau(x,y)\cdot O(\|\hat{s}_{t_k}(\hat{Y}_{t_k})\|).\end{aligned}$$

This implies that the first-order part does not exceed

$$\tau(x,y)\cdot O\left(\|\hat{s}_{t_k}(\hat{Y}_{t_k})\|\cdot\left(\|\hat{s}_{t_k}(\hat{Y}_{t_k})\| + d\rho(x,y)\right)\right).$$

Now, for the second-order part, we can also calculate it in the local normal coordinates:

$$\begin{aligned}g^{\alpha\beta}\partial_{\alpha\beta}\tau(x,y) &= g^{\alpha\beta}\tau(x,y)\left(\partial_\alpha\int_0^1\langle\dot{x}_s,\tilde{s}_k(x_s)\rangle\mathrm{d}s\right)\left(\partial_\beta\int_0^1\langle\dot{x}_s,\tilde{s}_k(x_s)\rangle\mathrm{d}s\right)\\ &\quad + g^{\alpha\beta}\tau(x,y)\int_0^1\Big[(\partial_\beta g_{\alpha\gamma}+\partial_\alpha g_{\beta\gamma})\tilde{s}_k(x_s)^\gamma + s(g_{\alpha\gamma}\partial_\beta + g_{\beta\gamma}\partial_\alpha)\tilde{s}_k(x_s)^\gamma\\ &\qquad + s^2 x^\xi(\partial_\beta g_{\xi\gamma}\cdot\partial_\alpha + \partial_\alpha g_{\xi\gamma}\cdot\partial_\beta + \partial_{\alpha\beta}g_{\xi\gamma} + g_{\xi\gamma}\partial_{\alpha\beta})\tilde{s}_k(x_s)^\gamma\Big]\mathrm{d}s\\ &= \tau(x,y)\cdot O\Big(\|\hat{s}_{t_k}(\hat{Y}_{t_k})\|^2 + \|\hat{s}_{t_k}(\hat{Y}_{t_k})\|\left(d\rho(x,y)+\omega^{-1}+d\omega\right)\\ &\quad + d\|\hat{s}_{t_k}(\hat{Y}_{t_k})\|\rho(x,y)\left((\omega^{-1}+d\omega)\rho(x,y)d + d + \omega^{-2} + (d\omega)^2\right)\Big)\\ &= \tau(x,y)\cdot O\left(\|\hat{s}_{t_k}(\hat{Y}_{t_k})\|^2 + \|\hat{s}_{t_k}(\hat{Y}_{t_k})\|d^2\rho(x,y)\right).\end{aligned}$$

Bound on $|u_1(x,y)|$. Combining the above analysis, we obtain that, for $\rho(x,y)\le O(d^{-1/2})$, we have

$$|u_1(x,y)| \le \tau(x,y)\cdot O\left(d^2 + \|\hat{s}_{t_k}(\hat{Y}_{t_k})\|^2 + \|\hat{s}_{t_k}(Y_{t_k})\|d^{1.5}\right).$$

Bound on $\|\nabla_x u_1(x,y)\|$. Similarly as before, Lemma E.5 shows that, for $\rho(x,y)\le O(d^{-1/2})$,

$$\sum_{\alpha=1}^d\sum_{\beta=1}^d g^{\alpha\beta} = O(d),\quad \partial_\gamma g_{\alpha\beta} = O(d^{-1/2}),\quad \partial_{\gamma\eta}g_{\alpha\beta} = O(1).$$

Therefore, when take an additional derivative, when taking derivative on the term of form $g^{\alpha\beta}$, the sum of these elements is bounded by $\sum_{\alpha=1}^{d}\sum_{\beta=1}^{d}\partial_\gamma g^{\alpha\beta} = O(d^{3/2})$, and when taking derivative on the term of form $\partial_\gamma g_{\alpha\beta}$, the bound for each entry is also multiplied by $O(d^{1/2})$. Moreover, by Lemma E.6, differentiating terms of form $\partial_{\eta\gamma} g_{\alpha\beta}$ only multiplies the bound by a constant, when we assume Assumption 5.1. Moreover, when differentiating the term $\tau(x,y)$ ouside the integral, it introduces an additional $\frac{\partial\tau}{\tau} = O\left(\|\hat{s}_{t_k}(\hat{Y}_{t_k})\|\right)$ term to the bound. We skip the detailed calculation here. By applying the product rule on all the terms, we eventually obtain the following bound:

$$\|\nabla u_1(x,y)\| \le \tau(x,y)\cdot O\left[\left(d^2 + \|\hat{s}_{t_k}(\hat{Y}_{t_k})\|^2 + \|\hat{s}_{t_k}(Y_{t_k})\|d^{1.5}\right)\cdot\left(\|\hat{s}_{t_k}(\hat{Y}_{t_k})\| + d^{0.5}\right)\right].$$

Bound on $|\Delta u_1(x,y)|$. This part is almost the the same as calculating the gradient. However, taking the Laplacian sums over the partial derivatives, where the partial derivative $\partial_{\alpha\beta}$ is considered with weight $g^{\alpha\beta}$. Since $\sum_{\alpha=1}^{d}\sum_{\beta=1}^{d} g^{\alpha\beta} = O(d)$, this would introduce an additional $O(d)$ term into the bound. We also skip the detailed calculation for clarity, and eventually obtain the following bound:

$$|\Delta u_1(x,y)| \le \tau(x,y)\cdot O\left[d\left(d^2 + \|\hat{s}_{t_k}(\hat{Y}_{t_k})\|^2 + \|\hat{s}_{t_k}(Y_{t_k})\|d^{1.5}\right)\left(\|\hat{s}_{t_k}(\hat{Y}_{t_k})\| + d^{0.5}\right)^2\right].$$

Summarize. According to Assumption 5.2, $\|\hat{s}_{t_k}\| \le \frac{CD}{T-t_k} + \frac{C\sqrt{d}}{\sqrt{T-t_k}}$. Define $\delta_k := T - t_k$ for simplicity. Then by applying the assumed bound on the score function, we deduce that, as long as $\rho(x,y) \le \frac{c}{\sqrt{d}}$ for some universal constant c, we have

$$\begin{aligned}
|u_1(x,y)| &\le \tau(x,y)\cdot O\left(d^2\delta_k^{-1/2} + D^2\delta_k^{-2} + d\delta_k^{-1}\right),\\
\|\nabla u_1(x,y)\| &\le \tau(x,y)\cdot O\left(d^{5/2}\delta_k^{-1} + d^{3/2}\delta_k^{-1.5} + D^2 d^{1/2}\delta_k^{-5/2}\right),\\
|\Delta u_1(x,y)| &\le \tau(x,y)\cdot O\left(d^4\delta_k^{-3/2} + d^3\delta_k^{-2} + D^2 d^2\delta_k^{-3}\right).
\end{aligned}$$

D.3 Bound on the Infinite Series

We first provide upper bounds on the absolute value of

$$r_t(x,y) = (\partial_t + H)\tilde{k}_t^{(1)}(x,y) = (\partial_t + H)[\eta_\iota(\rho(x,y))\hat{q}_t(x,y)(u_0(x,y) + tu_1(x,y))].$$

For the remainder r_t, we partition them into two parts

$$r_t(x,y) = \hat{r}_t(x,y) + \tilde{r}_t(x,y),$$

where the first part $\hat{r}_t(x,y)$ only contains derivatives on $k_t^{(1)}(x,y)$, and the second part $\tilde{r}_t(x,y)$ takes at least one derivatives on $\eta_\iota(\rho(x,y))$.

For the first part, by that

$$(\partial_t + H)k_t^{(1)}(x,y) = t\hat{q}_t(x,y)[Bu_1(x,y)],$$

Hence,

$$|\hat{r}_t(x,y)| \le 1_{\rho(x,y)\le\iota} t\hat{q}_t(x,y)\,|Bu_1(x,y)|.$$

Recall that

$$\begin{aligned}
B = &-\frac{1}{2}\Delta + \langle\tilde{s}_k(x) - j^{1/2}\nabla j^{-1/2}, \nabla\rangle\\
&+ \operatorname{div}(\tilde{s}_k(x)) + \frac{1}{2}j^{1/2}\Delta j^{-1/2} + \langle V(x), j^{1/2}\nabla j^{-1/2}\rangle.
\end{aligned}$$

Then, apply the bounds from previous section and use Lemma E.5,

$$\begin{aligned}
|B_x u_1(x,y)| &\lesssim |\Delta u_1| + (\|\hat{s}_{t_k}(\hat{Y}_{t_k})\| + d\rho(x,y))\|\nabla u_1\| + O(\|\hat{s}_{t_k}(\hat{Y}_{t_k})\|d + d^2)|u_1|\\
&\lesssim \tau(x,y)\cdot\left(d^4\delta_k^{-3/2} + d^3\delta_k^{-2} + D^2 d\delta_k^{-3}\right)
\end{aligned}$$

For the second part, use that $H := -\frac{1}{2}\Delta + \langle \tilde{s}_k(x), \nabla \rangle + \mathrm{div}(\tilde{s}_k(x))$,

$$\begin{aligned}
|\tilde{r}_t(x,y)| &\le |k^1(x,y)| \cdot |\Delta\eta_\iota(\rho(x,y)) + \langle \tilde{s}_k(x), \nabla\eta_\iota(\rho(x,y))\rangle| + |\langle \nabla\eta_\iota(\rho(x,y)), \nabla k^1(x,y)\rangle| \\
&\overset{(a)}{\le} C\hat{q}_t(x,y)\tau(x,y)1_{\iota/2\le\rho(x,y)\le\iota}\left[\frac{1}{\iota^2} + \frac{1}{\iota}|\Delta\rho(x,y)| + \iota^{-1}(D\delta_k^{-1} + d^{1/2}\delta_k^{-1/2})\right. \\
&\qquad \left. + \iota^{-1}\left(\frac{|\nabla\rho(x,y)^2|}{2t} + j^{1/2}\nabla j^{-1/2} + \tau(x,y)^{-1}(|\nabla u_0| + t|\nabla u_1|)\right)\right] \\
&\overset{(b)}{\le} C\hat{q}_t(x,y)\tau(x,y)1_{\iota/2\le\rho(x,y)\le\iota}\left[\frac{d}{\iota^2} + \frac{D\delta_k^{-1} + d^{1/2}\delta_k^{-1/2}}{\iota}\right] \\
&\overset{(c)}{\le} C\hat{q}_t(x,y)\tau(x,y)1_{\iota/2\le\rho(x,y)\le\iota} \cdot \frac{1}{\iota t},
\end{aligned}$$

where we apply Lemma E.5 for (b) and assumed that $t \le \min\{\tau d^{-1}, \sup_{x\in B_\iota(y)}\{|u_1(x,y)|\}^{-1}\}$ for (a), (b) and (c).

In addition, as long as $\rho(x,y) \le c\left(D\delta_k^{-1} + d^{1/2}\delta_k^{-1/2}\right)$, by definition we know that

$$c \le \tau(x,y) \le C$$

for universal constants c and C, and that

$$\|g(x) - I\| \le \frac{1}{d}$$

holds when $\rho(x,y) \le \frac{c}{\sqrt{d}}$, so that $c \le j(x) \le C$ holds for universal constants. Therefore, Let $A = C \cdot \left(d^4\delta_k^{-3/2} + d^3\delta_k^{-2} + D^2 d\delta_k^{-3}\right)$, we can simply write

$$\begin{aligned}
|\hat{r}_t(x,y)| &\le q_t(x,y)1_{\rho(x,y)\le\iota} \cdot A \\
|\tilde{r}_t(x,y)| &\le q_t(x,y)1_{\iota/2\le\rho(x,y)\le\iota} \cdot \frac{1}{\iota t}.
\end{aligned}$$

Then, we are ready to control the infinite series. Focus on the term $Q^{(n)}(x,y)$. Following the steps by Lemma 24 by Xu et al. (2026), by taking $z_0 = y$, and $s_0 = 0$,

$$\begin{aligned}
\left|Q^{(k)}(t,x,y)\right| &\le A^n \int_{t\Delta^n} \mathrm{d}s_1 \cdots \mathrm{d}s_n \int_{\mathcal{M}^n} q_{s_i}(x,z_n)1_{\rho(z_{i-1},z_i)\le\iota} \\
&\qquad \cdot \prod_{i=1}^n \hat{q}_{s_i - s_{i-1}}(z_i, z_{i-1})\left((s_i - s_{i-1})1_{\rho(z_{i-1},z_i)\le\iota}\right. \\
&\qquad \left. + \frac{1}{(s_i - s_{i-1})\iota}1_{\iota/2\le\rho(z_{i-1},z_i)\le\iota}\tilde{P}\right)\mathrm{d}z_1 \cdots \mathrm{d}z_n \\
&\le A^n \int_{t\Delta^n} (t^n I_{\text{loc}}(s) + I_{\text{out}}(s))ds,
\end{aligned}$$

where I_{loc} is the integration over

$$\mathcal{R} = \{(z_1, \ldots, z_n) : \rho(z_i, y) \le 2\iota; \rho(z_i, z_{i-1}) \le \iota/2\},$$

and I_{out} performs integration over $\mathcal{R}^c$.

Now, we directly use the results from Xu et al. (2026). First they have showed that as long as $\iota \le \frac{1}{Cd^3}$,

$$I_{\text{loc}}(s) \le 8^{k+1} \cdot q_t(x,y), \quad I_{\text{out}}(s) \le 32^k (ct)^{-d/2} \exp\left(-\frac{\iota^2}{16td}\right).$$

Then, As long as $\rho(x,y) \le \iota/4d$, and $t \le \frac{c\iota^2}{d\log(d/\iota^2)}$, we have

$$I_{\text{out}}(s) \le C \cdot 32^k q_t(x,y)t.$$

This finally implies that

$$\left|Q^{(k)}(t,x,y)\right| \lesssim \frac{t^{2k}A^k 8^k + 32^k A^k t^{k+1}}{k!}.$$

Therefore, when $t \le (80A)^{-1}$, summing over k finally implies

$$\sum_{k=1}^{\infty}\left|Q^{(k)}(t,x,y)\right| \le t^2 A = o(t).$$

D.4 Put All Things Together

First, since the previous density bound are only satisfied in the local coordinate chart, we want to localize both sampling process $\tilde{Y}$ and the estimated reverse process $\hat{Y}$. Specifically, for each sampling step $\tilde{P}_i$, define an auxiliary process so that, as long as the sampled vector $\tilde{s}_k(\tilde{Y}_{t_i'}) + \sqrt{t_{i+1}' - t_i'}w_{i'} > \iota/(4d)$, re-sample a new vector, and repeat this process until the requirement is satisfied. Define its transition density as $\hat{Q}_{\text{loc}}(t,x,y)$. Moreover, for the estimated process $\hat{Y}_t$ whose transition density is $P(t,x,y)$, define

$$\hat{P}_{\text{loc}}(t,x,y) = \frac{\hat{P}(t,x,y)}{\int_M 1_{\rho(x,y)\le \iota/(4d)}\hat{P}(t,x,y)\mathrm{d}x},$$

that is, we perform the similar re-sample process. Clearly, the TV distance between sampling process and $\hat{Q}$ and between P and $\hat{P}$ is exactly the probability of re-sampling. Use the similar bound derived from BDG inequality as in Appendix B.4, this probability is bounded by $O(d^3(h_i')^3/(\iota/d)^6) = O(d^9(h_i')^3\iota^{-6}) = o((h_i')^2)$, where h_i' represents the step size for Brownian motion simulation.

Moreover, we can directly write the transition density $\hat{Q}_{\text{loc}}(t,x,y)$ as

$$\hat{Q}_{\text{loc}}(t,x,y) = \frac{1_{\rho(x,y)\le \iota/4d}}{P(\text{resample})}Q(t,x,y),$$

where

$$Q(t,x,y) := \frac{1}{(2\pi t)^{d/2}}\exp\left(-\frac{|\log_y x - \tilde{s}_k(y)\cdot t|^2}{2t}\right)j(x)^{-1}$$

and the dot product is done in Euclidean space.

D.4.1 Analysing the Density Ratio

Now, for $\rho(x,y) \le \iota/4d$, we provide an upper bound for

$$\begin{aligned}\left|\frac{\hat{P}(t,x,y)}{Q(t,x,y)} - 1\right| \le\ & C\left|\frac{\hat{P}(t,x,y)}{\hat{q}_t(x,y)u_0(x,y)} - 1\right| \\ & + C\left|\frac{\hat{q}_t(x,y)}{q_t(x,y)j(x)^{-1}} - 1\right| \\ & + C\left|\frac{q_t(x,y)u_0(x,y)j(x)^{-1}}{Q(t,x,y)} - 1\right|.\end{aligned}$$

By the previous analyses, the first term is bounded by $C\left(d^2\delta_k^{-0.5} + D^2\delta_k^{-2} + d\delta_k^{-1}\right)t$ for small t.

For the second term, it equals to $j(x)^{-1/2}$, and then the difference is upper bounded by

$$|j(x)^{-1/2} - 1| \le (1 + C\rho(x,y)^2)^{d/2} - 1 \le Cd\rho(x,y)^2,$$

since $\rho(x,y) \le \frac{c}{\sqrt{d}}$.

For the third term, directly expanding the term, we get

$$\begin{aligned}&\left|\frac{q_t(x,y)u_0(x,y)j(x)^{-1}}{Q(t,x,y)}-1\right|\\=&\left|\exp\left(\frac{|\log_y x-t\cdot\tilde{s}_k(y)|^2}{2t}-\frac{\rho(x,y)^2}{2t}-\int_0^1\langle\dot{x}_s,\tilde{s}_k(x_s)\rangle_g ds\right)-1\right|\\=&\left|\exp\left(t|\tilde{s}_k(y)|^2+\int_0^1 x^\alpha(\delta_{\alpha\beta}\tilde{s}_k(y)^\beta+g_{\alpha\beta}\tilde{s}_k(x_s)^\beta)\mathrm{d}s\right)-1\right|.\end{aligned}$$

Clearly, when $t<\left(\frac{D^2}{\delta_k^2}+\frac{d}{\delta_k}\right)^{-1}$,

$$\exp\left(t|\tilde{s}_k(y)|\right)=1+O(D^2\delta_k^{-2}+d\delta_k^{-1})\cdot t.$$

Moreover, $|g(x)-I|\le C\rho(x,y)^2$, and by the definition of $\tilde{s}_k$,

$$|\nabla\tilde{s}_k(x)|\le\|\hat{s}_{t_k}(\hat{Y}_{t_k})\|(\omega^{-1}+\sqrt{d}\max\Gamma)\le\sqrt{d}\cdot O(D\delta_k^{-1}+d^{1/2}\delta_k^{-1/2}).$$

So,

$$\left|\int_0^1 x^\alpha(\delta_{\alpha\beta}\tilde{s}_k(y)^\beta+g_{\alpha\beta}\tilde{s}_k(x_s)^\beta)\mathrm{d}s\right|\le\rho(x,y)^3\|\tilde{s}_k(x_s)\|+\max|\nabla\tilde{s}_k(x)|\rho(x,y)^2.$$

Therefore, for $\rho(x,y)\le\iota/(4d)$, combining Assumption 5.2 and our previous analyses on $\nabla\tilde{s}_k$,

$$\begin{aligned}\left|\frac{\hat{P}(t,x,y)}{Q(t,x,y)}-1\right|\lesssim&(d\delta_k^{-1/2}+Dd^{1/2}\delta_k^{-1})\rho(x,y)^2+(d^{1/2}\delta_k^{-1/2}+D\delta_k^{-1})\rho(x,y)^3\\&+(D^2\delta_k^{-2}+d\delta_k^{-1}+d^2\delta_k^{-1/2})\cdot t.\end{aligned}$$

D.5 Controlling the KL divergence

Now we compute $\mathrm{KL}(\hat{Q}_{\mathrm{loc}}\|\hat{P}_{\mathrm{loc}})$. by that $\log(1+x)\ge x-Cx^2$ for $x\in[-1/2,1/2]$,

$$\begin{aligned}\mathrm{KL}(\hat{Q}_{\mathrm{loc}}\|\hat{P}_{\mathrm{loc}})=&-\int_M\hat{Q}_{\mathrm{loc}}(t,x,y)\log\frac{\hat{P}_{\mathrm{loc}}(t,x,y)}{\hat{Q}_{\mathrm{loc}}(t,x,y)}\mathrm{d}x\\\le&-\int_{\mathcal{M}}\hat{Q}_{\mathrm{loc}}(t,x,y)\left[\frac{\hat{P}_{\mathrm{loc}}(t,x,y)}{\hat{Q}_{\mathrm{loc}}(t,x,y)}-1-C\left|\frac{\hat{P}_{\mathrm{loc}}(t,x,y)}{\hat{Q}_{\mathrm{loc}}(t,x,y)}-1\right|^2\right]\mathrm{d}x\\\le&C\int_{\mathcal{M}}\hat{Q}_{\mathrm{loc}}(t,x,y)\left|\frac{\hat{P}_{\mathrm{loc}}(t,x,y)}{\hat{Q}(t,x,y)}-1\right|^2\mathrm{d}x,\end{aligned}$$

where we have used that

$$\int_{\mathcal{M}}\hat{Q}(t,x,y)\mathrm{d}x=\int_{\mathcal{M}}\hat{P}(t,x,y)\mathrm{d}x=1.$$

Therefore, when t is sufficiently small,

$$\begin{aligned}\mathrm{KL}(\hat{Q}_{\mathrm{loc}}\|\hat{P}_{\mathrm{loc}})\le&C\mathbb{E}_{\hat{Q}_{\mathrm{loc}}}\left|\frac{\hat{P}(t,x,y)}{Q(t,x,y)}-1\right|^2+C\sup\left|\frac{\hat{Q}_{\mathrm{loc}}}{Q}-1\right|^2+C\sup\left|\frac{\hat{P}_{\mathrm{loc}}}{\hat{P}}-1\right|^2\\\le&C(d^2\delta_k^{-1}+D^2d\delta_k^{-2})\mathbb{E}_{\hat{Q}_{\mathrm{loc}}}[\rho(x,y)^4]\\&+C(d\delta_k^{-1}+D^2\delta_k^{-2})\mathbb{E}_{\hat{Q}_{\mathrm{loc}}}[\rho(x,y)^6]\\&+C(D^4\delta_k^{-4}+d^2\delta_k^{-2}+d^4\delta_k^{-1})t^2+o(t^4)\\\le&C\left(D^4\delta_k^{-4}+d^4\delta_k^{-1}+D^2\delta_k^{-2}d^3\right)t^2.\end{aligned}$$

Therefore, let $\tilde{Y}_{\text{loc}}$ and $\hat{Y}_{\text{loc}}$ be two processes defined using localized transition density $\hat{Q}_{\text{loc}}$ and $\hat{P}_{\text{loc}}$, respectively. Then, given that the step sizes satisfy $h'_i := t'_{i+1} - t'_i \le h'\delta_k$ for a scale parameter $h' > 0$, the KL divergence between the distributions of $\tilde{Y}_{\text{loc},T-\delta}$ and $\hat{Y}_{\text{loc},T-\delta}$ is bounded by

$$\begin{aligned}\mathrm{KL}\left(\mathrm{Law}(\tilde{Y}_{\text{loc},T-\delta})\,\middle\|\,\mathrm{Law}(\hat{Y}_{\text{loc},T-\delta})\right) &\le h'\sum_{i'=0}^{N'-1} h_i\left(\frac{D^4}{(T-t'_i)^3} + d^4 + \frac{D^2d^3}{T-t'_i}\right)\\ &\le h'\left(d^4T + \delta^{-2}D^4 + D^2d^3\log(T/\delta)\right).\end{aligned}$$

Proof of Theorem 5.3 With Pinsker's inequality and the triangle inequality for TV distance,

$$\begin{aligned}\mathrm{TV}\left(\mathrm{Law}(\tilde{Y}_{T-\delta}), \mathrm{Law}(\hat{Y}_{T-\delta})\right)^2 &\lesssim \sqrt{\mathrm{KL}\left(\mathrm{Law}(\tilde{Y}_{\text{loc},T-\delta})\,\middle\|\,\mathrm{Law}(\hat{Y}_{\text{loc},T-\delta})\right)}\\ &\quad + \mathrm{TV}\left(\mathrm{Law}(\hat{Y}_{\text{loc},T-\delta}), \mathrm{Law}(\hat{Y}_{T-\delta})\right)^2\\ &\quad + \mathrm{TV}\left(\mathrm{Law}(\tilde{Y}_{T-\delta}), \mathrm{Law}(\tilde{Y}_{\text{loc},T-\delta})\right)^2\\ &\lesssim h'\left(d^4T + \delta^{-2}D^4 + D^2d^3\log(T/\delta)\right).\end{aligned}$$

D.6 Proof of Corollary 5.4

proof of Corollary 5.4. Using Pinsker's inequality, triangle inequality for TV distance, and the data-processing inequality,

$$\mathrm{TV}(p_\delta, \tilde{q}_{T-\delta}) \lesssim \mathrm{TV}(p_T, \hat{q}_0) + \sqrt{\mathrm{KL}(p_\delta, \hat{q}^*_{T-\delta})} + \mathrm{TV}(\hat{q}_{T-\delta}, \tilde{q}_{T-\delta}),$$

where $\hat{q}^*$ refer to the distribution defined by equation (7), but replace the initial distribution $\hat{q}_0$ by p_T instead. Then, apply Theorem 4.3 on the KL divergence and apply Theorem 5.3 on the last TV distance, as long as the step sizes h and h' are sufficiently small,

$$\mathrm{TV}(p_\delta, \tilde{q}_{T-\delta}) \lesssim \mathrm{TV}(p_T, \hat{q}_0) + \epsilon_{\text{score}} + \sqrt{hd\log(T/\delta)} + \sqrt{h'd^4T}.$$

Therefore, by choosing $h^{-1} = \Theta(\epsilon^{-2}d\log(T/\delta))$ and $(h')^{-1} = \Theta\left(\epsilon^{-2}d^4T\right)$, we can achieve $O(\epsilon)$ TV error. Finally, we conclude using the relations $N = \Theta\left(h^{-1}\log(T/\delta)\right)$ and $N' = \Theta\left((h')^{-1}\log(T/\delta)\right)$, which comes from our definition of step size schedule. □

E Useful Lemmas

First, we provide some upper bounds on the score functions $\nabla \log p_t(x)$. Although the results can be found in many literatures (e.g., Engoulatov (2006); Hsu (1999)), the dimension d are often dealt as a constant and hided in the formulas. Therefore, we first summarize a useful results in the following lemma, where the dependence on d are calculated explicitly.

Let $H(t, x, y)$ be the transition kernel of standard Brownian motion on $\mathcal{M}$.

Lemma E.1. *Suppose the Ricci curvature is bounded below by $-C_R$, where $C_R \ge 0$. Then,*

$$\|\nabla \log H(t, x, y)\| \le C\sqrt{\left(\frac{1}{t} + C_R\right)\left(d + \frac{\rho(x,y)^2}{t} + \sqrt{C_R d}\rho(x,y)\right)}.$$

Proof. This lemma rewrites the results by Engoulatov (2006), where the only difference is that we write the d-related constants explicitly. First, corollary 6 by Engoulatov (2006) shows that

$$\|\nabla \log H(t, x, y)\|^2 \le 2\left(\frac{1}{t} + C_R\right)\mathbb{E}\log\frac{H(t/2, X_{t/2}, y)}{H(t, x, y)}.$$

Moreover, according to Equation (22) from Engoulatov (2006), for all $s > 0, t > 0$ and $x, y \in \mathcal{M}$,

$$\frac{H(t, x)}{H(t+s, y)} \le \left(\frac{t+s}{t}\right)^{n/2} \times \exp\left(\frac{(d(x,y) + \sqrt{C_R d}s)^2}{4s} + \frac{\sqrt{C_R d}}{4}\min\{\rho(x,y), \sqrt{C_R d}s\}\right).$$

Replace t and s by $t/2$, we obtain

$$\begin{aligned}\mathbb{E}\log\frac{H(t/2,X_{t/2},y)}{H(t,x,y)} &\le \frac{n}{2}\log 2+\frac{(\rho(x,y)+\sqrt{C_Rd}t/2)^2}{2t}\\ &\quad+\frac{\sqrt{C_Rd}}{4}\min\{\rho(x,y),\sqrt{C_Rd}t/2\}\\ &= C_1\left(d+\frac{\rho(x,y)^2}{t}+\sqrt{C_Rd}\rho(x,y)\right),\end{aligned}$$

where C_1 is a universal constant.

Hence,

$$\begin{aligned}\|\nabla\log H(t,x,y)\| &\le \sqrt{2\left(\frac{1}{t}+R\right)\mathbb{E}\log\frac{H(t/2,X_{t/2},y)}{H(t,x,y)}}\\ &\le C\sqrt{\left(\frac{1}{t}+C_R\right)\left(d+\frac{\rho(x,y)^2}{t}+\sqrt{C_Rd}\rho(x,y)\right)}\end{aligned}$$

□

The next lemma is a direct application of the former lemma, and is widely used in our analysis. First, recall the forward process is the standard brownian motion on manifold $\mathcal{M}$ with initial distribution p_0.

Lemma E.2. *Under Assumption 4.1, for all $t>0$ and $x\in\mathcal{M}$, the score function $\nabla\log p_t(x)$ satisfies*

$$\|\nabla\log p_t(x)\|\le C\left(\frac{\sqrt{d}}{\sqrt{t}}+\frac{D}{t}\right).$$

Moreover, for all $t>0$,

$$\mathbb{E}_{x\sim p_t}\left[\|\nabla\log p_t(x)\|^2\right]\le\frac{Cd}{t}.$$

Proof. For $0\le s<t$, we can represent

$$p_t(x)=\int_{\mathcal{M}}p_s(\mathrm{d}y)H(t-s,x,y).$$

Then,

$$\begin{aligned}\nabla\log p_t(x) &= \frac{\nabla p_t(x)}{p_t(x)}\\ &= \frac{\int_{\mathcal{M}}p_0(\mathrm{d}y)\nabla_x H(t,x,y)}{\int_M p_0(\mathrm{d}y)H(t,x,y)}\\ &= \frac{\int_{\mathcal{M}}p_0(\mathrm{d}y)H(t,x,y)\nabla_x\log H(t,x,y)}{\int_{\mathcal{M}}p_0(\mathrm{d}y)H(t,x,y)}\\ &= \mathbb{E}\left[\nabla_x\log H(t,x,X_0)|X_t=x\right].\end{aligned}$$

By Lemma E.2, since Assumption 4.1 ensures that the Ricci curvature is nonnegative, we can use $C_R=0$ in Lemma E.1 to deduce that

$$\|\nabla\log H(t,x,y)\|\le C\left(\frac{\sqrt{d}}{\sqrt{t}}+\frac{\rho(x,y)}{t}\right).$$

Hence, by that $\rho(x,y)\le D$, we directly obtain a uniform upper-bound, which is the first claim in Lemma E.2:

$$\|\nabla\log p_t(x)\|\le\sup_{y\in\mathcal{M}}\|\nabla_x\log H(t,x,y)\|\le C\left(\frac{\sqrt{d}}{\sqrt{t}}+\frac{D}{t}\right).$$

In addition, taking expectation over $x \sim p_t$, we have

$$\begin{aligned}\mathbb{E}_{x\sim p_t}\left[\|\nabla \log p_t(x)\|^2\right] &= \mathbb{E}_{x\sim p_t}\left[\left\|\mathbb{E}_{y\sim p_{0|t}}[\nabla_x \log H(t,x,y)]\right\|^2\right] \\ &\overset{\text{(a)}}{\leq} \mathbb{E}_{x\sim p_t}\mathbb{E}_{y\sim p_{0|t}}\left[\|\nabla_x \log H(t,x,y)\|^2\right] \\ &\leq C\cdot \mathbb{E}_{(y,x)\sim p_{0,t}}\left[\frac{d}{t}+\frac{\rho(x,y)^2}{t^2}\right] \\ &\overset{\text{(b)}}{=} C\cdot\frac{d}{t}+\frac{C}{t^2}\cdot\mathbb{E}\left[\rho(X_t,X_0)^2\right].\end{aligned}$$

Here, step (a) comes from Jensen's inequality applied on convex function $\|\cdot\|^2$, and step (b) is by noticing that $(y,x)\sim p_{0,t}$ have the same distribution as X_0, X_t, and recall that the forward process $(X_t)_{t\geq 0}$ is a standard Brownian motion on $\mathcal{M}$.

We now upper-bound the expectation of $\rho(X_t,X_0)^2$. By fixing X_0 and take $r(x)=\rho(x,X_0)$, the function $r(\cdot)$ is smooth on $\mathcal{M}\setminus(\text{Cut}(X_0)\cap\{X_0\})$, where $\text{Cut}(X_0)$ represents the cut locus of X_0 on $\mathcal{M}$. Moreover, the Laplacian comparison theorem (see, e.g., Theorem 3.4.2 of Hsu 2002) shows that, when the Ricci curvature is nonnegative,

$$\Delta r(x)\leq\frac{d-1}{r(x)},\quad x\in\mathcal{M}\setminus(\text{Cut}(X_0)\cap\{X_0\}).$$

With this in mind, by applying Ito's lemma on $r(X_t)$, we obtain that

$$\begin{aligned}\mathrm{d}r^2(X_t) &= \langle\nabla r^2(X_t),\mathrm{d}X_t\rangle+\frac{1}{2}\Delta r^2(X_t)\mathrm{d}t \\ &= 2r(X_t)\cdot\mathrm{d}\beta_t+\frac{1}{2}(2r(X_t)\Delta r(X_t)+2\|\nabla r(X_t)\|^2)\mathrm{d}t \\ &\leq 2r(X_t)\cdot\mathrm{d}\beta_t+\left(\frac{(d-1)r(X_t)}{r(X_t)}+1\right)\mathrm{d}t \\ &= 2r(X_t)\cdot\mathrm{d}\beta_t+d\cdot\mathrm{d}t,\end{aligned}$$

where β_t is a 1-dimensional standard Brownian motion in $\mathbb{R}$. Note that $r^2(x)$ is smooth at X_0, so the formula is valid until X_t hits the cut locus. However, we can apply the Calabi's trick on the cut locus to show that the formula holds globally. Specifically, let $x_0\in\text{Cut}(X_0)$, we pick a minimizing geodesic $\gamma:[0,r(x)]\to\mathcal{M}$ from X_0 to x_0, and then define $r_\epsilon(x)=\epsilon+\rho(\gamma(\epsilon),x)$. Clearly, $r_\epsilon(x_0)=r(x_0)$ and $r_\epsilon\geq r$ on $\mathcal{M}$. Moreover, Calabi's trick ensures that r_ϵ is smooth around x_0. In addition, $\Delta r_\epsilon^2(x)\leq 2d$ also holds for all $\epsilon>0$. Now, suppose $X_s\in\text{Cut}(X_0)$, by applying Ito's formula on r_ϵ, we obtain that locally,

$$r^2(X_t)-r^2(X_s)\leq r_\epsilon^2(X_t)-r_\epsilon^2(X_s)=\int_s^t 2\rho(\gamma_\epsilon,X_r)\mathrm{d}\tilde{\beta}_r+d(t-s),$$

Taking $\epsilon\to 0$, and taking summation of all local hits on the cut locus, we show that

$$r^2(X_t)\leq\int_0^t 2r(X_s)d\beta_s+d\cdot t$$

holds for all $t>0$. Therefore, by taking expectation,

$$\mathbb{E}[r^2(X_t)]\leq dt,$$

which immediately implies that

$$\begin{aligned}\mathbb{E}_{x\sim p_t}\left[\|\nabla\log p_t(x)\|^2\right] &\leq C\cdot\frac{d}{t}+\frac{C}{t^2}\cdot\mathbb{E}\left[\rho(X_t,X_0)^2\right] \\ &\leq\frac{Cd}{t}.\end{aligned}$$

This concludes the second claim of Lemma E.2. □

The second inequality of Lemma E.2 provides an upper bound for the second moment of $\nabla\log p_t(x)$. The following lemma further computes a bound for the higher moments.

Lemma E.3. *Under Assumption 4.1, for all $t > 0$,*

$$\mathbb{E}_{x\sim p_t}\left[\|\nabla \log p_t(x)\|^4\right] \le \frac{Cd^2}{t^2}.$$

More generally, for any $p > 1$,

$$\mathbb{E}_{x\sim p_t}\left[\|\nabla \log p_t(x)\|^{2p}\right] \le \frac{C_p d^p}{t^p},$$

where the constant C_p only depends on p.

Proof. Similarly as the proof of Lemma E.2, using Jensen's inequality,

$$\mathbb{E}_{x\sim p_t}\left[\|\nabla \log p_t(x)\|^{2p}\right] \le \frac{C_p d^p}{t^p} + \frac{C_p}{t^{2p}} \cdot \mathbb{E}\left[\rho(X_t, X_0)^{2p}\right].$$

Let $r(x) := \rho(x, X_0)$. We have proved that

$$r^2(X_t) \le \int_0^t r(X_s)\mathrm{d}\beta_s + d \cdot t.$$

Using $(x^p)' = px^{p-1}$ and $(x^p)'' = p(p-1)x^{p-2}$ apply Itô's lemma on the right hand side, we have

$$r^{2p}(X_t) \le \int_0^t pr^{2p-2}(X_s) \cdot [r(X_s)\mathrm{d}\beta_s + d \cdot \mathrm{d}s] + \frac{1}{2}\int_0^t p(p-1)r^{2(p-2)}(X_s)r(X_s)^2\mathrm{d}s$$

Taking expectation, we obtain that

$$\begin{aligned}
\mathbb{E}[r^{2p}(X_t)] &\le \int_0^t \mathbb{E}[r^{2p-2}(X_s)] \cdot C_p d\mathrm{d}s \\
&\le \int_0^t \left(\mathbb{E}[r^{2p-2}(X_s)]/t^{(p-1)/p}\right) \cdot C_p d t^{(p-1)/p}\mathrm{d}s \\
\text{(Young's inequality)} \quad &\le \int_0^t \left\{\frac{\mathbb{E}[r^{2p-2}(X_s)]^{p/(p-1)}}{tp/(p-1)} + \frac{C_p d^p t^{p-1}}{p}\right\}\mathrm{d}s \\
&\le \int_0^t \frac{p-1}{tp}\mathbb{E}[r^{2p}(X_s)]\mathrm{d}s + C_p d^p t^p.
\end{aligned}$$

By Grönwall's inequality,

$$\mathbb{E}[r^{2p}(X_t)] \le \exp\left(\int_0^t \frac{p-1}{tp}\mathrm{d}s\right) C_p d^p t^p \le C_p d^p t^p.$$

This concludes the desired results. □

In addition, we provide a bound to compare with the second inequality in Lemma E.2. In the next lemma, we do not assume that the Ricci curvature to be nonnegative, but instead, we assume that the Ricci curvature is bounded below by $-C_R$ for $C_R \ge 0$. When $C_R = 0$, it immediately reduces to the nonnegative Ricci curvature case.

Lemma E.4. *Under Assumption 4.1, but instead we assume that the Ricci curvature is bounded below by $-C_R$. Then,*

$$\mathbb{E}_{x\sim p_t}\|\nabla \log p_t(x)\|^2 \le C\left(\frac{1}{t} + C_R\right)\left(d + \sqrt{dC_R}D\right).$$

Proof. Similarly as before, we still need to compute $\mathbb{E}[r(X_t)^2]$ for $r(x) := \rho(X_0, X_t)$. However, in this case, Laplacian comparison theorem indicates that

$$\Delta r \le (d-1)k\coth(kr),$$

where $k := \sqrt{\frac{C_R}{d-1}}$. Thus,

$$\frac{1}{2}\Delta r^2 \le \|\nabla r\|^2 + r\Delta r \le 1 + (d-1)kr\coth(kr) \le 1 + (d-1)(1+kr) = d + (d-1)kr,$$

where we have used the fact that $x\coth x \le 1 + x$ for $x \ge 0$. Therefore, by applying the same analysis as before, we can obtain that

$$\begin{aligned}\mathbb{E}[r^2(X_t)] &\le \int_0^t \mathbb{E}\left[d + (d-1)kr(X_s)\right] \mathrm{d}s \\ &\le \left(d + \sqrt{(d-1)C_R}D\right) t\end{aligned}$$

Finally, applying Lemma E.1, and apply the same analysis with Jensen's inequality,

$$\mathbb{E}_{x\sim p_t}\|\nabla \log p_t(x)\|^2 \le C\left(\frac{1}{t} + C_R\right)\left(d + \sqrt{(d-1)C_R}D\right)$$

□

The next lemma by Xu et al. (2026) is very useful, which quantifies the local behavior of Riemannian metric matrix.

Lemma E.5 (Lemma 5 by Xu et al. (2026))**.** *Let $y \in M$. Under Assumption 4.1, the Riemannian metric matrix $g(x)$ under the local normal coordinates at y satisfies*

$$\begin{aligned}\|g(y) - I\| &\le C \cdot C_{Rm}\rho(x,y)^2, \\ \|\partial_\alpha g(y)\| &\le C \cdot C_{Rm}\rho(x,y), \\ \|\partial_{\alpha\beta} g(y)\| &\le C \cdot C_{Rm},\end{aligned}$$

as long as $\rho(x,y) \le c/\sqrt{K}$. Here, $C > 0$ is a universal constant, and $c > 0$ only depends on r_i.

Remark: In the original paper (Xu et al., 2026), the authors said that c, C are polynomial in d. However, in their proof, the only property required was that $K\rho(x,y)^2$ is not too large, and the constants are actually universal.

In addition, under Assumption 5.1, we can provide upper bounds for third and fourth derivatives.

Lemma E.6. *Under Assumption 4.1 and 5.1, in the same setting of Lemma E.5, we further have*

$$\begin{aligned}\|\partial_{\alpha\beta\gamma} g(y)\| &\le C, \\ \|\partial_{\alpha\beta\gamma\eta} g(y)\| &\le C,\end{aligned}$$

where C only depends on C_{Rm} and C'_{Rm}.

Proof. In the original proof of Lemma E.5 (Lemma 5 in Xu et al. 2026), they define a matrix $\boldsymbol{J}$ along the geodesic curve γ from y to x so that $g(x) = \boldsymbol{J}(1)^T\boldsymbol{J}(1)$ and the matrix Jacobi equation

$$\boldsymbol{J}''(s) = -\boldsymbol{R}(s)\boldsymbol{J}(s), \quad J(0) = 0, \quad J'(0) = I.$$

is satisfied. Here the matrix $\boldsymbol{R}(s)$ is defined by

$$\left(\boldsymbol{R}(s)\boldsymbol{u}\right)_\alpha := \left\langle R\left(u^\beta e_\beta(s), \dot\gamma(s)\right)\dot\gamma(s), e_\alpha(s)\right\rangle,$$

where R represents the Riemmanian curvature tensor, and $\{e_\beta\}_{\beta=1}^n$ is an orthonormal frame obtained from the parallel transport along γ.

In the proof of Lemma E.5, Xu et al. (2026) have already shown that $\|\boldsymbol{J}\| \le C\,\|\partial_\alpha \boldsymbol{J}\| \le C\rho(x,y)$, and $\|\partial_{\alpha\beta}\boldsymbol{J}\| \le C$, where the derivatives are taken with respect to the endpoint x. Now, differentiate the matrix Jacobi equation to the third order, we obtain that

$$\partial_{\alpha\beta\gamma}\boldsymbol{J}''(s) = -\partial_{\alpha\beta\gamma}(\boldsymbol{R}(s)\boldsymbol{J}(s)).$$

Integrating twice with respect to s, and apply Assumption 5.1 to control the derivative of matrix $\boldsymbol{R}$, we can obtain

$$\partial_{\alpha\beta\gamma}\boldsymbol{J}(s) = \int_0^s (s-\tau)\left[C\partial_{\alpha\beta\gamma}\boldsymbol{J}(\tau) + C\cdot(C\rho(x,y)^2 + C\rho(x,y) + C)\right] d\tau.$$

Hence, by Grönwall's inequality, we eventually obtain that $\|\partial_{\alpha\beta\gamma}\boldsymbol{J}(1)\| \leq C$ whenever $\rho(x,y) \leq \frac{c}{\sqrt{d}}$. Therefore, apply the product rule on $g(x) = \boldsymbol{J}(1)^T\boldsymbol{J}(1)$,

$$\|\partial_{\alpha\beta\gamma}g(x)\| \leq C \cdot C + C \cdot C\rho(x,y) \leq C.$$

The case for the fourth derivative is exactly the same, and the key is again that the constant in Assumption 5.1 does not depend on the dimension d. □