POLYNOMIAL CONVERGENCE OF RIEMANNIAN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models have demonstrated remarkable empirical success in the recent years and are considered one of the state-of-the-art generative models in modern AI. These models consist of a forward process, which gradually diffuses the data distribution to a noise distribution spanning the whole space, and a backward process, which inverts this transformation to recover the data distribution from noise. Most of the existing literature assumes that the underlying space is Euclidean. However, in many practical applications, the data are constrained to lie on a submanifold of Euclidean space. Addressing this setting, De Bortoli et al. (2022) introduced Riemannian diffusion models and proved that using an exponentially small step size yields a small sampling error in the Wasserstein distance, provided the data distribution is smooth and strictly positive, and the score estimate is L_∞-accurate. In this paper, we greatly strengthen this theory by establishing that, under L_2-accurate score estimate, a *polynomially small stepsize* suffices to guarantee small sampling error in the total variation distance, without requiring smoothness or positivity of the data distribution. Our analysis only requires mild and standard curvature assumptions on the underlying manifold. The main ingredients in our analysis are Li-Yau estimate for the log-gradient of heat kernel, and Minakshisundaram-Pleijel parametrix expansion of the perturbed heat equation. Our approach opens the door to a sharper analysis of diffusion models on non-Euclidean spaces.

1 INTRODUCTION

Initially introduced by Sohl-Dickstein et al. (2015) and later advanced by Song & Ermon (2019); Ho et al. (2020); Dhariwal & Nichol (2021), diffusion model has become one of the bedrocks in generative modeling across a variety of application domains such as vision, video, speech, and many others. On a high level, diffusion generative models operate by coupling two stochastic processes:

1. A *forward* process
$$X_0 \xrightarrow{\text{add noise}} X_1 \xrightarrow{\text{add noise}} \cdots \xrightarrow{\text{add noise}} X_T,$$
where X_0 is sampled from the target distribution p_0 in $\mathbb{R}^d$, and X_T resembles pure noise.
2. A *reverse* process
$$Y_T \xrightarrow{\text{denoise}} Y_{T-1} \xrightarrow{\text{denoise}} \cdots \xrightarrow{\text{denoise}} Y_0,$$
where Y_T starts from pure noise, and gradually removes the noise, so that at Y_0, we recover a new sample from a distribution close to p_0.

Diffusion models build a reverse process to recover the target data distribution by step-wise reversing the forward process, with a goal of matching the probabilities $Y_t \approx X_t$ in distribution for $t \in \{T, \ldots, 1\}$. Leveraging the theory of backward stochastic differential equations (SDE) (Anderson, 1982; Haussmann & Pardoux, 1986), this can be formally achieved as soon as the score functions – the gradients of the log marginal density of the forward process – become available, which can be estimated via score matching (Hyvärinen & Dayan, 2005).

Tremendous recent progresses have been made in understanding the convergence of diffusion models in the Euclidean space, e.g. Lee et al. (2023b); Chen et al. (2023); Benton et al. (2024); Li et al.

(2024), which establish near-tight polynomial iteration complexities of discrete-time samplers under L_2-accurate score estimates and mild assumptions of the data distribution. These convergence results provide strong justifications to the empirical success of diffusion models for generating from complex multi-modal distributions.

Many scientific domains, however, are intrinsically *non-Euclidean*; examples include orientations on $\mathrm{SO}(3)$, directions on spheres, toroidal angles, articulated poses, and symmetric positive definite (SPD) matrices are naturally modeled on Riemannian manifolds (Piggott & Solo, 2016; Muniz et al., 2022). Recently, there has been an increasing interest in effectively sampling from distributions supported on manifolds and providing theoretical guarantees (Girolami & Calderhead, 2011; Gatmiry & Vempala, 2022; Li & Erdogdu, 2023; Guan et al., 2025). Although sampling on manifolds has been studied extensively (Cheng et al., 2023), extending diffusion models to manifolds requires careful treatments to incorporate the manifold constraints into both the time-inhomogeneous forward and reverse processes, with selected attempts in De Bortoli et al. (2022); Huang et al. (2022); Lou et al. (2023); Liu et al. (2023); Fishman et al. (2023).

One notable development is De Bortoli et al. (2022), who introduced *Riemannian Score-Based Generative Models* (RSGMs) with convergence guarantees in the Wasserstein distance. Specifically, they established a time-reversal diffusion process for geometric Brownian motion on manifolds, which can be similarly learned via score matching (Hyvärinen & Dayan, 2005). While groundbreaking, their convergence bound suffers from a few caveats: 1) it requires an exponentially small stepsize, leading to a possibly exponential iteration complexity in some of the manifold parameters; 2) it requires L_∞-accurate score estimates, which are impractical in deep learning; and 3) the data distribution is required to be smooth and strictly positive on compact manifolds. This naturally raises the following questions:

Can we achieve polynomial iteration complexity for manifold diffusion models using L_2-accurate score estimates under mild data assumptions?

1.1 Our contribution

We provide a discrete-time analysis of the RSGM sampler in De Bortoli et al. (2022), assuming L_2-accurate score estimates. Under mild geometric conditions of the manifold without assuming smooth or strictly positive data densities, we establish that *polynomial* stepsizes suffice for accurate sampling on manifolds in *total variation* (TV). More precisely, for some $\varepsilon > 0$, the TV error between the distribution of the output Y_0 and p_δ, that is, an approximation to p_0 with early-stopping time $\delta > 0$, obeys

$$\mathsf{TV}\big(p_\delta, \mathsf{Law}(Y_0)\big) \lesssim \varepsilon + \varepsilon_{\mathsf{score}} + \sqrt{hT}\,\mathrm{poly}(d, \delta^{-1}),$$

as long as the horizon satisfies $T \gtrsim \lambda_1^{-1}(d \log d + \log(d/\varepsilon))$. Here, d is the dimension of $\mathcal{M}$, λ_1 is the spectral gap of $-\Delta_M$, $h > 0$ is the stepsize, and $\varepsilon_{\mathsf{score}}$ is the L_2 score estimation error.[1] This bound suggests that a polynomial stepsize is sufficient: take $T \asymp \lambda^{-1}(d \log d + \log(d/\varepsilon))$ and $h = \frac{\varepsilon^2}{\mathrm{poly}(d,\delta^{-1})T}$, then the TV error is bounded by $\varepsilon + \varepsilon_{\mathsf{score}}$ after an iteration complexity of $N = T/h \asymp \mathrm{poly}(d, \delta^{-1})/(\lambda_1\varepsilon)^2$. This conveys a much more benign message about the efficiency of Riemannian diffusion models, compared with the iteration complexity in De Bortoli et al. (2022) that scales exponentially with the dimension d, under relaxed assumptions on both the data distribution and the score estimates.

Techniques. Our proof couples three ingredients: (i) high-probability Li-Yau gradient bounds for the manifold heat kernel together with early stopping to control $\|\nabla \log p_t\|$ without assuming positivity/smoothness of p_0; (ii) a localization scheme that "freezes" drifts across nearby tangent spaces but preserves continuous Brownian motion (BM), to separate the effects of discretizing scores and BM; and (iii) a quantitative estimates for Minakshisundaram–Pleijel parametrix that controls one-step deviations between the manifold heat flow and its discretized proxy. These components allow us to decompose the global TV error into (a) mixing, (b) score approximation, and (c) discretization remainder terms, each handled carefully to avoid exponential dependence.

[1] For simplicity, we omitted polynomial dependence on manifold geometry parameters and poly-log dependence on ε. The complete version can be found in Theorem 1.

Work	Structure	Metric	Iteration complexity	Data distribution
Benton et al. (2024)	Euclidean	TV	$\widetilde{O}\big(d/\varepsilon^2\big)$	bounded moment
Li et al. (2024)	Euclidean	TV	$\widetilde{O}(\mathrm{poly}(d)/\varepsilon)$	bounded support
Li & Yan (2025)	Euclidean	TV	$\widetilde{O}(d/\varepsilon)$	bounded moment
De Bortoli et al. (2022)	Manifold	W_p	$\widetilde{O}\big(\exp(O(d))/\varepsilon^{-1/\lambda_1}\big)$[a]	smooth, strictly positive
This work	Manifold	TV	$\tilde{O}\big(\frac{\mathrm{poly}(d)}{\lambda_1^2\varepsilon^2}\big)$	None (early stopping)

[a]The original W_p error in De Bortoli et al. (2022) is stated in the form of $\widetilde{O}\big(C\mathrm{e}^{-\lambda_1 T} + \mathrm{e}^T\sqrt{h}\big)$, where C is defined in Proposition C.6 therein, which in turn is specified by Urakawa (2006, Proposition 2.6) as the supremum of $t^{d/2}H(t,x,y)$, where H is the heat kernel on the manifold. In general, the best estimate for this is due to Li-Yau (Li & Yau, 1986), which gives $C \le \mathrm{e}^{O(d)}$. To achieve ε-error, we must set $T = \lambda_1^{-1}(\Omega(d) + \log\varepsilon^{-1})$ and $h = \mathrm{e}^{-2T}\varepsilon^2$, then the iteration complexity T/h has the claimed form.

Table 1: Comparison of the current theoretical guarantees on diffusion probabilistic models on Euclidean spaces and manifolds. Here, $\lambda_1 > 0$ is the spectral gap of the Laplace-Beltrami operator.

1.2 Related works

Non-asymptotic convergence for Euclidean diffusion models. Early convergence analyses of diffusion models require L_∞-accurate score estimates (De Bortoli et al., 2021). For stochastic samplers such as DDPM (Ho et al., 2020), early bounds under Lipschitz/smoothness assumptions of the data distribution admit an $O(T^{-\frac{1}{2}})$ iteration complexity in the total variation distance assuming L_2-accurate score estimates (Chen et al., 2023), with subsequent analyses relaxing the Lipschitz assumption yet retaining the same complexity (Lee et al., 2023a; Benton et al., 2024; Li et al., 2024). More recently, Li & Yan (2025) has improved the iteration complexity to $\widetilde{O}(T^{-1})$. For deterministic samplers, Chen et al. (2023) established polynomial convergence with exact scores, and Li et al. (2024) established a convergence rate of $O(T^{-1})$ under L_2-accurate scores. See Beyler & Bach (2025); Liang et al. (2024); Li & Jiao (2024) for additional analyses that established convergence in the Wasserstein distance and improved discrete-time rates. Several works (Li & Yan, 2024; Liang et al., 2025; Huang et al., 2024; Potaptchik et al., 2024) also developed non-asymptotic convergence rates of diffusion models under the manifold hypothesis, suggesting diffusion models are adaptive to low-dimensional structures. This line of work should not be confused with ours, where the diffusion process is designed specifically to be constrained on the manifold.

Sampling on Riemannian manifold. Cheng et al. (2022; 2023) analyzed the geometric Euler–Maruyama (EM) discretization for time-homogeneous SDEs, and proved a polynomial complexity guarantee under dissipative-distant geometric assumptions on the manifold. See also Bharath et al. (2025) for follow-ups. Guan et al. (2025) proposed a Riemannian proximal sampler with convergence guarantees under the log-Sobolev inequality. Various sampling algorithms are also studied for a related problem known as sampling from constrained spaces (Srinivasan et al., 2024; Ahn & Chewi, 2021). Nonetheless, convergence analyses of Riemannian diffusion models under general data distributions remain highly limited, with De Bortoli et al. (2022) being the only prior work with non-asymptotic convergence rates.

2 Backgrounds

2.1 Diffusion Models on Euclidean Space

We briefly recall diffusion processes on $\mathbb{R}^d$. Let $(W_t)_{t\ge 0}$ be a standard Brownian motion in $\mathbb{R}^d$.

Forward SDE and Fokker–Planck. Given a drift term $b_t : \mathbb{R}^d \to \mathbb{R}^d$, the forward noising process $(X_t)_{t\in[0,T]}$ solves the Itô SDE

$$\mathrm{d}X_t \;=\; b_t(X_t)\,\mathrm{d}t \;+\; \mathrm{d}W_t, \qquad X_0 \sim p_0.$$

Let p_t denote the law of X_t, then p_t satisfies the Fokker–Planck equation $\partial_t p_t = -\nabla(b_t p_t) + \frac{1}{2}\Delta p_t$. In the standard forward diffusion used in score-based diffusion models, $p_t = p_0 * \varphi_t$ is a Gaussian smoothing of p_0 with kernel $\varphi_t(z) = (2\pi t)^{-d/2} \exp\big(-\|z\|^2/(2t)\big)$.

Score and Reverse Process. The *score* of the forward process at time t is defined as $s_t(x) := \nabla \log p_t(x)$. The time-reversal identity (Anderson, 1982) yields a reverse-time process $(Y_t)_{t\in[0,T]}$ whose marginals match those of $(X_t)_{t\in[0,T]}$ which solves the reverse-time SDE (note that t flows from T to 0):

$$dY_t \;=\; [-b_t(Y_t) + s_t(Y_t)]\,\mathrm{d}t + \mathrm{d}W_t, \qquad Y_T \sim p_T.$$

In the driftless setting where $b_t \equiv 0$, this simplifies to

$$dY_t = s_t(Y_t)\mathrm{d}t + \mathrm{d}W_t.$$

Discretization (Euler–Maruyama). Let $0 = t_0 < t_1 < \cdots < t_N = T$, where $t_i - t_{i-1} =: h$ be a time grid. Writing q_k as the law of the reversed process Y_k at time t_k, the Euler–Maruyama process is given by

$$y_{k-1} = y_k + h\,[-b_{t_k}(y_k) + s_\theta(t_k, y_k)] + \sqrt{h}g_k, \quad g_k \sim \mathcal{N}(0, I_d).$$

In our driftless setting, this reduces to

$$y_{k-1} = y_k + hs_\theta(t_k, y_k) + \sqrt{h}g_k, \quad g_k \sim \mathcal{N}(0, I_d).$$

2.2 Geometry and notation

We assume some familiarity with Riemannian geometry, and make use of standard notation. Please refer to Jost (2017); Petersen (2006) for a more in-depth treatment. In particular, we use $\alpha, \beta, \xi, \zeta$, etc., to index *coordinate representation* of tensors, and assume Einstein's summation convention. Let $(\mathcal{M}, g)$ be a connected, compact d-dimensional Riemannian manifold, with geodesic distance $d(\cdot,\cdot)$ and volume measure μ. We assume $\mu(\mathcal{M}) = 1$. The Levi–Civita connection is denoted by ∇, and the Laplace–Beltrami operator by

$$\Delta_{\mathcal{M}} f := \nabla_\alpha \nabla^\alpha f.$$

We use $T_x\mathcal{M}$ for the tangent space at x and use $\exp_x : T_x\mathcal{M} \to \mathcal{M}$ for the exponential map and $\log_x$ for its local inverse on the normal neighborhood of x. The injectivity radius of $\mathcal{M}$ is assumed to be lower bounded by some $\rho \in (0, 1)$. Furthermore, we define *geodesic diameter* of (M, g) is

$$\mathrm{Diam}(M) \;:=\; \sup_{x,y\in M} d(x, y),$$

where $d(\cdot,\cdot)$ is the geodesic distance induced by g. We further denote Rm as the Riemannian curvature tensor. Geodesic ball centered at x with radius r is denoted $B_x(r)$.

2.3 Heat flow, Brownian motion, and diffusion on $\mathcal{M}$

We also recall the setup for SDE and diffusion processes on Riemannian manifolds introduced in De Bortoli et al. (2022); Cheng et al. (2023). Let $(W_t)_{t\geq 0}$ be a standard Brownian motion in $\mathbb{R}^d$ and $U_x : \mathbb{R}^d \to T_x\mathcal{M}$ any orthonormal frame at x. The *Geometric Brownian motion* solves

$$\mathrm{d}X_t = U_{X_t} \circ \mathrm{d}W_t,$$

where $\circ$ denotes Stratonovich integral, and its transition density $p_t(x, y)$ with respect to μ solves the heat equation

$$\partial_t p_t(\cdot, y) = \frac{1}{2}\Delta_{\mathcal{M}} p_t(\cdot, y).$$

Equivalently, Brownian motion can be defined abstractly as the solution to the martingale problem for the operator $\frac{1}{2}\Delta_{\mathcal{M}}$. Concretely, for any $f \in C^\infty([0,\infty) \times \mathcal{M})$, the process

$$M_t^f := f(t, X_t) - f(0, X_0) - \int_0^t \left(\partial_s + \frac{1}{2}\Delta_{\mathcal{M}}\right) f(s, X_s)\mathrm{d}s$$

Algorithm 1 Riemannian Score-Based Generative Models (RSGM)

1: Manifold $(\mathcal{M}, g)$; score $s_\theta(x,t)$; early stopping time $\delta > 0$; reverse time grid $\delta = t_0 < t_1 < \cdots < t_N = T$; step size $h = t_k - t_{k-1}$; initial $x_N \sim \mu$ (uniform distribution);
2: **for** $k \in \{N, \dots, 1, 0\}$ **do**
3: Choose an orthonormal frame U_k at Y_k, which is a linear map from $\mathbb{R}^d$ to $T_{Y_k}\mathcal{M}$.
4: $g_k \sim \mathcal{N}(0, I_d)$ in $\mathbb{R}^d$; $\quad G_k \leftarrow U_k g_k \in T_{Y_k}\mathcal{M}$.
5: $b_k \leftarrow s_\theta(Y_k, t_k) \in T_{Y_k}\mathcal{M}$
6: $\Delta_k \leftarrow h b_k + \sqrt{h}\, G_k \in T_{Y_k}\mathcal{M}$
7: **if** $\|\Delta_k\| \le h^{1/4}$ **then**
8: $Y_{k-1} \leftarrow \exp_{Y_k}(\Delta_k)$
9: **else**
10: $Y_{k-1} \sim \mu$
11: **return** Y_0

is a martingale with respect to the natural filtration of X. More generally, a forward diffusion process with drift is given by

$$\mathrm{d}X_t \;=\; b_t(X_t)\,\mathrm{d}t + U_{X_t} \circ \mathrm{d}W_t,$$

with Fokker–Planck equation $\partial_t p_t = -\nabla(b_t p_t) + \frac{1}{2}\Delta_{\mathcal{M}} p_t$. Note that in this setting, the following process is a martingale for smooth f:

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

Let p_t denote the density of X_t w.r.t. μ, and define the *score* $s_t := \nabla \log p_t$. The time-reversal identity on manifolds yields a reverse SDE:

$$\mathrm{d}\widetilde{X}_t = (-b_t(\widetilde{X}_t) + \nabla \log p_t(\widetilde{X}_t))\mathrm{d}t + U_{\widetilde{X}_t} \circ \mathrm{d}W_t,$$

In our algorithm, the score $\nabla \log p_t$ is approximated by a trained neural network $s_\theta(t, x)$.

On compact manifolds, $-\Delta_{\mathcal{M}}$ admits a spectral gap $\lambda_1 > 0$. Any initial distribution mixes to the uniform distribution μ along the heat flow with rate $\mathrm{e}^{-\lambda_1 t}$. We use the *total variation* (TV) and the Kullback-Leibler distance to measure two distributions p, q:

$$\mathsf{TV}(p,q) = \int_{\mathcal{M}} \big|\mathrm{d}p - \mathrm{d}q\big|, \quad \mathsf{KL}(p \,\|\, q) = \int_{\mathcal{M}} \left(\log \frac{\mathrm{d}p}{\mathrm{d}q} \right) \mathrm{d}p.$$

3 Main Result

In this section, for completeness, we first introduce the RSGM algorithm in De Bortoli et al. (2022) with an injective radius guardrail. Then, we offer a polynomial convergence guarantee in Theorem 1. Recall that the forward diffusion on a Riemannian manifold $(\mathcal{M}, g)$ is

$$\mathrm{d}X_t \;=\; U_{X_t} \circ \mathrm{d}W_t, \qquad X_0 \sim p_0.$$

The time-reversal identity yields the *reverse-time SDE*

$$\mathrm{d}Y_t \;=\; s_t(Y_t)\,\mathrm{d}t \;+\; U_{Y_t} \circ \mathrm{d}W_t, \qquad Y_T \sim p_T. \tag{2}$$

In Algorithm 1, we provide the outline of discretized diffusion process on Riemannian manifold. In each reverse step $k \in \{N, \dots, 1, 0\}$, we select an orthonormal frame U_k at Y_k to map Euclidean noise to the tangent space at Y_k. At line 3, we sample Gaussian noise g_k and lift it to the tangent space $T_{Y_k}\mathcal{M}$, obtaining $G_k \in T_{Y_k}\mathcal{M}$. In line 5-6, we compute the reverse drift $b_k = s_\theta(Y_k, t_k)$ and propose a tangent update $\Delta_k = hb_k + \sqrt{h}G_k$. If Δ_k falls within the injective radius, we proceed to project it back to $\mathcal{M}$ with the exponential map $\exp_{Y_k}$. Otherwise, it triggers a reset $Y_{k-1} \sim \mu$. The algorithm terminates at $k = 0$ and returns the final iterate Y_0. Therefore, we ensure every update is well-defined in normal coordinates during the diffusion process while preserving a simple mixture kernel. We first formalize the assumptions needed for the convergence guarantee.

Assumption 1 (Regularity)**.** *Let $(\mathcal{M}, g)$ be a connected, compact d-dimensional Riemannian manifold. We assume the following conditions on $\mathcal{M}$:*

(A1) ***Positive injectivity radius:*** *there exists some $\rho \in (0,1)$ such that the injective radius $\geq \rho$.*

(A2) ***Uniform curvature bounds:*** *there exist constants $K \geq 1$ such that*

$$\max\left\{ \mathrm{Diam}(\mathcal{M}), \|\mathrm{Rm}\|_{L^\infty}, \|\nabla \mathrm{Rm}\|_{L^\infty}, \|\nabla^2 \mathrm{Rm}\|_{L^\infty} \right\} \leq K.$$

(A3) ***Regularity of score estimates:*** *there exists a polynomial $\mathrm{poly}(d, K, \rho^{-1})$, such that*

$$\|s_\theta(t_k, x)\| \leq \mathrm{poly}(d, K, \rho^{-1}) \left(\|\nabla \log p_{t_k}(x)\| + t_k^{-1} \mathrm{Diam}^2(\mathcal{M}) \right), \quad \forall x \in \mathcal{M}.$$

In Assumption 1, we made the standard "bounded geometry" assumption; similar assumptions also occur in Cheng et al. (2022); De Bortoli et al. (2022). A positive injective radius ρ ensures that we have sufficient room to operate on the tangent spaces as a proxy of operating on manifolds, since for every $x \in \mathcal{M}$, the exponential map $\exp_x$ is a diffeomorphism on the geodesic ball of radius ρ. Bounds on Riemannian tensors rule out pathological cases, which helps to control the error propagation along the reverse diffusion. Lastly, compactness ensures a positive spectral gap of $\Delta_{\mathcal{M}}$ with $\lambda_1 > 0$, which is necessary to guarantee that the forward process mixes. The mild assumption (A3) on the score estimates avoids excessively large drifts in diffusion, and can be implemented easily in practice by clipping. In addition to the above, we also need a standard assumption on the score estimation error (Chen et al., 2023).

Assumption 2 (Score estimation error)**.** *There exists $\varepsilon_{\mathsf{score}} > 0$ such that*

$$\sum_{k=1}^{N} (t_k - t_{k-1}) \mathbb{E} \|s_\theta(t_k, Y_{t_k}) - \nabla \log p_{t_k}(Y_{t_k})\|^2 \leq \varepsilon_{\mathsf{score}}^2.$$

With the above assumptions, we are now ready to present our main quantitative guarantee of RSGM, as outlined in the following TV-accuracy bound.

Theorem 1. *Assume Assumptions 1 and 2 hold. There exists some universal constant $C, C' > 0$ such that the following holds. If $T \geq \frac{C}{\lambda_1}(d \log(d/\rho) + K + \log(\frac{N}{\varepsilon}))$, then the output Y_0 of Algorithm 1 obeys*

$$\mathsf{TV}(p_\delta, \mathsf{Law}(Y_0)) \leq \varepsilon + C' \varepsilon_{\mathsf{score}} + \sqrt{hT}\, \mathrm{poly}(d, K, \rho^{-1}, \delta^{-1}),$$

where h is the discretization step size, $\lambda_1 > 0$ is the mixing rate of the geometric Brownian Motion on $\mathcal{M}$, i.e., the smallest eigenvalue of $-\Delta_{\mathcal{M}}$ in $L^2(\mu)$.

The bound decomposes cleanly into three terms: ε results from mixing of the heat semigroup at the spectral gap λ_1, $\varepsilon_{\mathsf{score}}$ captures error from imperfect score estimation, and $\sqrt{hT}\, \mathrm{poly}(d, K, \rho^{-1}, \delta^{-1})$ is the discretization error controlled by the step size and curvature. Consequently, choosing $T \asymp \lambda_1^{-1}(d \log d + \log(d/\varepsilon))$ and $h = \frac{\varepsilon^2}{\mathrm{poly}(d, K, \rho^{-1}\delta^{-1})T}$, then the TV error is bounded by $\varepsilon + \varepsilon_{\mathsf{score}}$ after polynomial iterations $N = T/h \asymp \mathrm{poly}(d, K, \rho^{-1}, \delta^{-1})/(\lambda_1 \varepsilon)^2$.

Compared to prior convergence rates in the Wasserstein metric (De Bortoli et al., 2022), which require exponential complexity, we achieve polynomial convergence of Riemannian diffusion models for the first time. Nonetheless, we emphasize that TV and Wasserstein distances are incomparable with each other in general, and our guarantee complements prior Wasserstein results (De Bortoli et al., 2022) by ensuring distributional closeness in a different notion with a much smaller number of iterations.

We note that the bound established in Theorem 1 holds under very mild geometric assumptions, requiring only constraints on the injective radius and Riemannian curvature. The purpose of this study is to demonstrate that, in the manifold setting, the exponential blow-up in T can be avoided and polynomial complexity can be achieved. To keep the exposition as simple as possible and to clearly highlight the key ideas, we have not attempted to optimize the current bound on the degree of the polynomial. Potential approaches for sharper bounds include: (i) a better design of discretization schedule and a more careful computation of discretization error, such as those in Li & Jiao (2024), Benton et al. (2024) (notably, the dependence on δ might be improved to poly-logarithmic in this way); (ii) a tailored analysis for TV error that does not rely on Pinsker's inequality, like those in Li & Yan (2025), may also be extended to manifolds; (iii) a tighter version of our Minakshisundaram-Pleijel parametrix bound. We leave these improvements as future work.

4 Proof Outline

Throughout the proof, we assume that

$$h \leq \frac{1}{\text{poly}(d, K, \rho^{-1}, \delta^{-1})}, \tag{3}$$

since otherwise the bound in Theorem 1 would be trivial (recall that TV is always bounded by 2). We start by recalling the sequence considered in RSGM. Let $(Y_k)_{k\in\{0,\ldots,N\}}$ be given by $Y_N \sim \mu$ and for any $k \in \{0, \ldots, N-1\}$:

$$Y_{k-1} = \begin{cases} \exp_{Y_k}\left[hs_\theta(t_k, Y_k) + \sqrt{h}G_k\right], & \|hs_\theta(t_k, Y_k) + \sqrt{h}\,G_k\| \leq h^{1/4}, \\ \text{drawn from } \mu, & \text{otherwise.} \end{cases}$$

This defines a sequence of probability transition kernels $\widehat{\mathsf{K}}_{t_k,t_{k-1}}$. For simplicity, we denote this by $\widehat{\mathsf{K}}_k$. Let q_k be the law of Y_k. We have

$$q_0 = q_N \widehat{\mathsf{K}}_N \widehat{\mathsf{K}}_{N-1} \cdots \widehat{\mathsf{K}}_1.$$

Similarly, the probability transition kernel from time t_k to t_{k-1} in (2) is denoted by $\mathsf{K}_{t_k,t_{k-1}}$ or K_k in short. We have

$$p_0 = p_N \mathsf{K}_N \mathsf{K}_{N-1} \cdots \mathsf{K}_1.$$

Our goal would be to bound $\mathsf{TV}(p_0, q_0)$ as in Theorem 1.

Step I. Constructing auxiliary kernels via localization. The overarch of our proof is to decompose the total error into four components:

$$\text{error} = (\text{initialization error}) + (\text{score error}) + (\text{drift discretization error}) + (\text{BM simulation error}).$$

More concretely:

- **Initialization error** arises from initializing Y_N with μ instead of the true marginal p_N;
- **Score error** arises from imperfect score estimation;
- **Drift discretization** error arises from approximating the continuous-time drift $s_\theta(t, Y_t)$ by its "time-frozen" counterpart $s_\theta(t_k, Y_k)$;
- **Brownian motion (BM) simulation error** is a distinctive feature of the manifold setting. Unlike in Euclidean space – where the transition kernel of Brownian motion over $[t_k, t_{k-1}]$ is exactly Gaussian with variance $(t_k - t_{k-1})$ – the transition kernel of manifold-valued Brownian motion cannot be simulated exactly by any discrete-time process, even after time discretization. This inherent inexactness gives rise to the final error term.

The first two components are relatively easy to bound using well-established tools: mixing rate bounds of heat flow (Urakawa, 2006) and Girsanov transform (Chen et al., 2023). For the drift discretization error, recent techniques developed in the Euclidean setting (Benton et al., 2024) can also be adapted with modifications that account for the manifold curvature. However, the last component – the Brownian motion simulation error – represents the core challenge in the manifold setting, which fundamentally denies a direct extension of Euclidean analysis.

In view of this, we introduce an intermediate random process that separates the drift discretization error from the BM simulation error. Constructing such a process, however, involves additional technicality. In particular, the frozen drift $s_\theta(t_k, Y_k)$ is a vector in the tangent space $T_{Y_k}\mathcal{M}$, and is therefore only well-defined at the fixed point Y_k. This poses a compatibility issue: as Brownian motion evolves continuously on the manifold, it immediately departs from Y_k, rendering the frozen drift ill-defined. Careful geometric considerations are thus required to reconcile the piecewise-constant drift approximation with the intrinsic curvature of the manifold.

In our analysis, this is handled using localization by the construction of an auxiliary sequence of transition kernels $\mathsf{K}_k^{\mathsf{aux}}$. These kernels do not appear in the algorithm itself; they serve solely as an analytical tool to facilitate the proof. These kernels expose the behavior of the time-reverse SDE (2) when the estimated score s_θ is frozen to be a constant vector field in between discretization steps, meanwhile keeping the continuous Brownian motion.

Let $\eta : [0,\infty) \to [0,1]$ be a smooth cutoff function, i.e., η is decreasing, $\eta|_{[0,1]} \equiv 1$ and $\eta|_{[2,\infty)} \equiv 0$. Such a function can be chosen such that $|\eta'| + |\eta''| + |\eta'''| \le 100$.

Recall that ρ is the injective radius of $\mathcal{M}$, and K is the curvature bound. Define

$$\omega := \frac{c_\omega}{d^4} \min\left(\rho, \frac{1}{K}\right), \qquad \eta_\omega(x) = \eta\left(\frac{16\|x\|^2}{\omega^2}\right), \quad x \in \mathbb{R}^d, \tag{4}$$

where $c_\omega > 0$ is a small universal constant. We have $\eta_\omega|_{B(0,\omega/4)} \equiv 1$ and $\eta_\omega|_{\mathbb{R}^d \setminus B(0,\omega/2)} \equiv 0$. For $t > 0$, $x, y \in \mathcal{M}$, define the following vector field on $\mathcal{M}$:[2]

$$\mathscr{S}_{t,x}(y) = (\mathrm{d}\exp_x)_{\log_x y}\left(\eta_\omega(\log_x y) \cdot s_\theta(t,x)\right) \in T_y\mathcal{M}.$$

Intuitively speaking, $\mathscr{S}_{t,x}(\cdot)$ is the "constant" velocity field $s_\theta(t,x)$ in normal coordinates, which represents our idea of freezing the drift term for a time period. The $\mathrm{d}\exp_x$ in the formula is responsible for identifying $T_y\mathcal{M}$ with $T_x\mathcal{M}$. [3] On the other hand, the cut-off function η_ω is necessary to keep all our discussions restricted to the injective radius, so as to avoid pathologies of cut locus.

With this in mind, we are ready to define $\mathsf{K}_k^{\mathsf{aux}}$ as the transition kernel from time t_k to t_{k-1} of the reverse-time SDE

$$\mathrm{d}Y_t = \mathscr{S}_{t_k, Y_{t_k}}(Y_t)\mathrm{d}t + \mathrm{d}W_t, \quad t \in [t_{k-1}, t_k].$$

Then we define

$$p_k^{\mathsf{aux}} = p_N \mathsf{K}_N^{\mathsf{aux}} \mathsf{K}_{N-1}^{\mathsf{aux}} \cdots \mathsf{K}_{k+1}^{\mathsf{aux}}, \quad k = N, N-1, \cdots, 0.$$

Step II. Decomposing different sources of error. We now decompose

$$\mathsf{TV}(p_0, q_0) \le \mathsf{TV}(p_0, p_0^{\mathsf{aux}}) + \mathsf{TV}(p_0^{\mathsf{aux}}, q_0) \le \sqrt{2\mathsf{KL}(p_0 \parallel p_0^{\mathsf{aux}})} + \mathsf{TV}(p_0^{\mathsf{aux}}, q_0),$$

where the last inequality used Pinsker's inequality. To control $\mathsf{KL}(p_0 \parallel p_0^{\mathsf{aux}})$, we further introduce the counterpart of $\mathscr{S}_{t,x}$ using the exact score function $\nabla \log p_t$:

$$\mathscr{S}^\star_{t,x}(y) = (\mathrm{d}\exp_x)_{\log_x y}\left(\eta_\omega(\log_x y) \cdot \nabla \log p_t(x)\right) \in T_y\mathcal{M}.$$

A standard application of Girsanov (Chen et al., 2023; De Bortoli et al., 2022) then implies

$$\mathsf{KL}(p_0 \parallel p_0^{\mathsf{aux}}) \le \sum_{k=1}^N \int_{t_{k-1}}^{t_k} \mathbb{E}\left\|\nabla \log p_t(Y_t) - \mathscr{S}_{t_k, Y_{t_k}}(Y_t)\right\|^2 \mathrm{d}t$$

$$\le \underbrace{2\sum_{k=1}^N \int_{t_{k-1}}^{t_k} \mathbb{E}\|\nabla \log p_t(Y_t) - \mathscr{S}^\star_{t_k, Y_{t_k}}(Y_t)\|^2 \mathrm{d}t}_{\text{drift discretization}} + \underbrace{2\sum_{k=1}^N \int_{t_{k-1}}^{t_k} \mathbb{E}\|\mathscr{S}_{t_k, Y_{t_k}}(Y_t) - \mathscr{S}^\star_{t_k, Y_{t_k}}(Y_t)\|^2 \mathrm{d}t}_{\text{score matching}}. \tag{5}$$

It remains to decompose $\mathsf{TV}(p_0^{\mathsf{aux}}, q_0)$. To isolate the initialization error, we introduce

$$q_0^\star = p_N \widehat{\mathsf{K}}_N \widehat{\mathsf{K}}_{N-1} \cdots \widehat{\mathsf{K}}_1.$$

By triangle inequality and post-processing inequality, we have

$$\mathsf{TV}(p_0^{\mathsf{aux}}, q_0) \le \mathsf{TV}(p_0^{\mathsf{aux}}, q_0^\star) + \mathsf{TV}(q_0^\star, q_0) \le \underbrace{\mathsf{TV}(p_0^{\mathsf{aux}}, q_0^\star)}_{\text{BM simulation}} + \underbrace{\mathsf{TV}(p_N, q_N)}_{\text{initialization}}.$$

[2] More precisely, the formula is obviously well-defined for $d(x,y) < \omega$. On the boundary and outside of this radius, we can smoothly extend $\eta_\omega(\log_x y) = 0$ (even though $\log_x y$ might be undefined) and therefore $\mathscr{S}_{t,x}(y) = 0$. Thus the definition extends smoothly to the whole manifold.

[3] Generally speaking, it is more natural to use parallel transport to identify different tangent spaces. However, this would later lead to a much more complicated treatment of perturbed heat equation with variable drifts. We choose to use $\mathrm{d}\exp$ here for simplicity.

Step III. Controlling initialization and score errors. By our design, $q_N = \mu$, and $\mathsf{TV}(p_N, q_N) = \mathsf{TV}(p_N, \mu)$. This is known as the mixing rate of heat flow in total variation norm, and has well-established bounds (e.g., Urakawa (2006)). The score-matching error, on the other hand, can be controlled with an analysis on the distortion on the Riemannian metric in normal coordinates. We compile the bounds into the following lemma.

Lemma 1. *There exists a universal constant $C > 0$, such that whenever $T \geq 1$, we have*

$$\mathsf{TV}(p_N, q_N) \leq \mathrm{e}^{C(K+d\log d)}\mathrm{e}^{-\frac{\lambda_1}{2}(T-\frac{1}{2})},$$

$$\sum_{k=1}^{N}\int_{t_{k-1}}^{t_k} \mathbb{E}\|\mathscr{S}_{t_k,Y_{t_k}}(Y_t) - \mathscr{S}^\star_{t_k,Y_{t_k}}(Y_t)\|^2 \mathrm{d}t \leq 2\varepsilon^2_{\mathsf{score}}.$$

Step IV. Controlling drift discretization error with Itô/Stratonovich calculus and Li-Yau estimates. The drift discretization error defined in (5) has a similar form to the discretization error for Euclidean setting (Benton et al., 2024), though additional complication arises due to non-constant $\mathscr{S}^\star_{t_k,Y_{t_k}}$. The idea is to study the time derivative of $\mathbb{E}\|\nabla \log p_t(Y_t) - \mathscr{S}^\star_{t_k,Y_{t_k}}(Y_t)\|^2$, which in view of $\partial_t \log p_t = -\frac{1}{2}\Delta_{\mathcal{M}} p_t$ (negative sign due to reverse time) involves space derivatives of $\log p_t$ up to third order. Fortunately, after applying Itô/Stratonovich calculus to simplify the expression, a key property in the proof of the Euclidean setting carries over: third-order derivatives of $\log p_t$ cancel out. The remaining first and second-order derivatives can be controlled by Li-Yau estimates on the log-gradient of the heat equation. We obtain

Lemma 2. *Under the assumptions as in Theorem 1, there is a universal constant $C > 0$ such that*

$$\sum_{k=1}^{N}\int_{t_{k-1}}^{t_k} \mathbb{E}\|\nabla \log p_t(Y_t) - \mathscr{S}^\star_{t_k,Y_{t_k}}(Y_t)\|^2 \mathrm{d}t \leq \frac{Cd^6K^8}{\delta^3}h^2 N.$$

Step V. Controlling BM simulation error using parametrix estimates. Our approach is inspired by the following consequence of post-processing inequality and Pinsker's inequality:

$$\mathsf{TV}(p_0^{\mathsf{aux}}, q_0^\star) \leq \sqrt{2\mathsf{KL}(p_0^{\mathsf{aux}} \,\|\, q_0^\star)} \leq \sqrt{2\sum_{k=1}^{N}\mathsf{KL}(p_k^{\mathsf{aux}}\mathsf{K}_k^{\mathsf{aux}} \,\|\, p_k^{\mathsf{aux}}\widehat{\mathsf{K}}_k)}.$$

This leads us to compare the kernel $\mathsf{K}_k^{\mathsf{aux}}$ and $\widehat{\mathsf{K}}_k$. In normal coordinates, Fokker-Planck equation shows that these two are the solutions of the heat equations with the Euclidean Laplacian and with the manifold Laplace-Beltrami operator. We utilize the Minakshisundaram-Pleijel parametrix theory (Berline et al., 2003) in geometric analysis for this comparison, and established a quantitative bound in polynomially small radius and short-time.

5 Conclusion

We developed a discrete-time theory for Riemannian diffusion models showing that a polynomial stepsize suffices for TV-accurate sampling under mild geometric conditions. Our main bound decomposes the total variation error into (i) exponential mixing at the spectral gap of $-\Delta_{\mathcal{M}}$, (ii) a term driven by score-estimation error, and (iii) a discretization error that scales as $\sqrt{h}\,\mathrm{poly}(d, K, \rho^{-1}, \delta^{-1})$. In particular, choosing stepsize h polynomially small in manifold parameters achieves any prescribed TV target without exponential blow-ups in dimension or curvature. This bound gives an alternative to prior Wasserstein-type guarantees by providing a strong distributional closeness criterion aligned with generative modeling goals.

References

Kwangjun Ahn and Sinho Chewi. Efficient constrained sampling via the mirror-Langevin algorithm. *Advances in Neural Information Processing Systems*, 34:28405–28418, 2021.

Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. ISSN 0304-4149. doi: https://doi.org/10.1016/0304-4149(82)90051-5.

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

Nicole Berline, Ezra Getzler, and Michele Vergne. *Heat kernels and Dirac operators*. Springer Science & Business Media, 2003.

Eliot Beyler and Francis Bach. Convergence of deterministic and stochastic diffusion-model samplers: A simple analysis in Wasserstein distance. *arXiv preprint arXiv:2508.03210*, 2025.

Karthik Bharath, Alexander Lewis, Akash Sharma, and Michael V Tretyakov. Sampling and estimation on manifolds using the langevin diffusion. In *Advances in Neural Information Processing Systems*, 2025.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2023.

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

Xiang Cheng, Jingzhao Zhang, and Suvrit Sra. Theory and algorithms for diffusion processes on riemannian manifolds, 2023.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in neural information processing systems*, 34:17695–17709, 2021.

Valentin De Bortoli, Emile Mathieu, Michael John Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

Nic Fishman, Leo Klarner, Valentin De Bortoli, Emile Mathieu, and Michael Hutchinson. Diffusion models for constrained domains. *arXiv preprint arXiv:2304.05364*, 2023.

Khashayar Gatmiry and Santosh S. Vempala. Convergence of the riemannian langevin algorithm. *ArXiv*, abs/2204.10818, 2022.

Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73:123 – 214, 03 2011. doi: 10.1111/j.1467-9868.2010.00765.x.

Alfred Gray. *Tubes*, volume 221. Springer Science & Business Media, 2003.

Yunrui Guan, Krishnakumar Balasubramanian, and Shiqian Ma. Riemannian proximal sampler for high-accuracy sampling on manifolds. *arXiv preprint arXiv:2502.07265*, 2025.

Richard S Hamilton. Matrix harnack estimate for the heat equation. *Communications in analysis and geometry*, 1(1):113–126, 1993.

Qing Han and Qi S Zhang. An upper bound for hessian matrices of positive solutions of heat equations. *The Journal of Geometric Analysis*, 26(2):715–749, 2016.

Ulrich G. Haussmann and Étienne Pardoux. Time reversal of diffusions. *Annals of Probability*, 14: 1188–1205, 1986.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.

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

Zhihan Huang, Yuting Wei, and Yuxin Chen. Denoising diffusion probabilistic models are optimally adaptive to unknown low dimensionality. *arXiv preprint arXiv:2410.18784*, 2024.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Jürgen Jost. *Riemannian Geometry and Geometric Analysis*. Universitext. Springer, Cham, 7 edition, 2017. ISBN 978-3-319-61859-3. doi: 10.1007/978-3-319-61860-9.

Holden Lee, Jiaming Lu, and Yu Tan. Convergence of score-based generative modeling for general data distributions. In *Algorithmic Learning Theory (ALT)*, pp. 946–985, 2023a.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *Proceedings of the 34th International Conference on Algorithmic Learning Theory*, volume 201 of *PMLR*, pp. 946–985, 2023b.

Gen Li and Yuchen Jiao. Improved convergence rate for diffusion probabilistic models. In *The Thirteenth International Conference on Learning Representations*, 2024.

Gen Li and Yuling Yan. Adapting to unknown low-dimensional structures in score-based diffusion models. *Advances in Neural Information Processing Systems*, 37:126297–126331, 2024.

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, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards non-asymptotic convergence for diffusion-based generative models. In *International Conference on Learning Representations (ICLR)*, 2024.

Junfang Li and Xiangjin Xu. Differential harnack inequalities on riemannian manifolds i: Linear heat equation. *Advances in Mathematics*, 226(5):4456–4491, 2011. ISSN 0001-8708. doi: https://doi.org/10.1016/j.aim.2010.12.009.

Mufan Bill Li and Murat A. Erdogdu. Riemannian langevin algorithm for solving semidefinite programs. *Bernoulli*, 29:3093 – 3113, 2023.

Peter Li and Shing Yau. On the parabolic kernel of the schödinger operator. *Acta Mathematica*, 156: 153–201, 07 1986. doi: 10.1007/BF02399203.

Jiadong Liang, Zhihan Huang, and Yuxin Chen. Low-dimensional adaptation of diffusion models: Convergence in total variation. *arXiv preprint arXiv:2501.12982*, 2025.

Yang Liang, Peng Ju, Yingbin Liang, and Ness Shroff. Non-asymptotic convergence of discrete-time diffusion models: New approach and improved rate. *arXiv preprint arXiv:2402.13901*, 2024.

Guan-Horng Liu, Tianrong Chen, Evangelos Theodorou, and Molei Tao. Mirror diffusion models for constrained and watermarked generation. *Advances in Neural Information Processing Systems*, 36:42898–42917, 2023.

Aaron Lou, Minkai Xu, Adam Farris, and Stefano Ermon. Scaling Riemannian diffusion models. *Advances in Neural Information Processing Systems*, 36:80291–80305, 2023.

Michelle Muniz, Matthias Ehrhardt, Michael Günther, and Renate Winkler. Higher strong order methods for linear itô sdes on matrix lie groups. *BIT Numerical Mathematics*, 62, 01 2022. doi: 10.1007/s10543-021-00905-9.

Peter Petersen. *Riemannian Geometry*, volume 171 of *Graduate Texts in Mathematics*. Springer, New York, 2 edition, 2006. ISBN 978-0-387-29246-5. doi: 10.1007/978-0-387-29403-2.

M. Piggott and V. Solo. Geometric euler–maruyama schemes for stochastic differential equations in so(n) and se(n). *SIAM Journal on Numerical Analysis*, 54:2490–2516, 08 2016. doi: 10.1137/15M1019726.

Yury Polyanskiy and Yihong Wu. *Information theory: From coding to learning*. Cambridge university press, 2025.

Peter Potaptchik, Iskander Azangulov, and George Deligiannidis. Linear convergence of diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2410.09046*, 2024.

Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*, volume 293. Springer Science & Business Media, 2013.

Richard M. Schoen, , and Shing-Tung Yau. Lectures on differential geometry. *Conference proceedings and lecture notes in geometry and topology*, 1, 1994.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Neural Information Processing Systems*, 2019.

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

Hajime Urakawa. Convergence rates to equilibrium of the heat kernels on compact riemannian manifolds. *Indiana University mathematics journal*, pp. 259–288, 2006.

A PRELIMINARIES

We first introduce some tools we use in the rest of the proof.

Lemma 3 (Pinsker's inequality, Polyanskiy & Wu (2025))**.** *For any two probability distributions p, q on $\mathcal{M}$, we have*

$$\mathsf{TV}(p,q) \le \sqrt{2\mathsf{KL}(p \,\|\, q)}.$$

Lemma 4. *Let v be a vector field on $\mathcal{M}$. In a local coordinate on $\mathcal{M}$, we have*

$$\partial_\alpha v_\beta = (\nabla_\alpha v)^\beta - \Gamma^\beta_{\alpha\gamma} v_\gamma.$$

Here $\Gamma^\beta_{\alpha\gamma}$ is the Christoffel symbol, defined as

$$\Gamma^\beta_{\alpha\gamma} = \frac{1}{2} g^{\beta\delta}(\partial_\alpha g_{\gamma\delta} + \partial_\gamma g_{\alpha\delta} - \partial_\delta g_{\alpha\gamma}).$$

Lemma 5 (Bounded Christoffel in normal coordinates)**.** *There exist universal constants $c, C > 0$ such that the following holds. Let $x \in \mathcal{M}$. In the normal coordinates (∂_α) at x, for any $y \in \mathcal{M}$ such that $d(x,y) \le c\min(\rho, 1/K)$, we have*

$$\begin{aligned} \|g(y) - I\| &\le CKd(x,y), \\ \|\partial_\alpha g_{\beta\gamma}\| &\le CKd(x,y), \\ \|\partial_{\alpha\beta} g_{\gamma\xi}\| &\le CK. \end{aligned}$$

Proof. This follows from the well-known Taylor expansion of g in normal coordinates (cf. Berline et al. (2003, Proposition 1.28)):

$$g_{\alpha\beta}(\exp_x(u)) = \delta_{\alpha\beta} - \frac{1}{3} R_{\alpha\gamma\beta\xi}(x) u^\gamma u^\xi + O\big((\|\mathrm{Rm}\| + \|\nabla \mathrm{Rm}\|)\|u\|^3\big), \quad \|u\| \le c\min(\rho, 1/K).$$

We sketch a formal proof here for sake of convenience. Denote $f_\theta(t) = \exp_x(t\theta)$, $\theta \in T_x\mathcal{M}$, which is the constant-speed geodesic starting from x with initial velocity θ. Consider the Jacobi field $E_\alpha(t,\theta) = (\mathrm{d}\exp_x)_{t\theta}(e_\alpha)$, which obeys the Jacobi equation

$$\frac{D^2}{\mathrm{d}t^2}E_\alpha(t,\theta) + \mathrm{Rm}(E_\alpha(t,\theta),\dot f_\theta)\dot f_\theta = 0, \quad E_\alpha(0,\theta) = 0, \quad \frac{D}{\mathrm{d}t}E_\alpha(0,\theta) = e_\alpha,$$

where, by convention, $\frac{D}{\mathrm{d}t}$ denotes covariant derivative. Applying our curvature bound assumption, we obtain

$$\left\|\frac{D^2}{\mathrm{d}t^2}E_\alpha(t,\theta)\right\| \le K\|\theta\|^2\|E_\alpha(t,\theta)\|.$$

Move everything to the base point x by parallel transport, we are left with an ODE inequality. Use Rauch comparison for this inequality to get a bound for $\|\tau_{\exp_x(t\theta)\to x}E_\alpha(t,\theta) - te_\alpha\|$, where τ is the parallel transport map. Note that parallel transport preserves inner product, thus

$$g_{\alpha\beta}(y) = \langle E_\alpha(1,\log_x y), E_\beta(1,\log_x y)\rangle = \langle \tau_{y\to x}E_\alpha(1,\log_x y), \tau_{y\to x}E_\beta(1,\log_x y)\rangle.$$

The desired bound on $\|g(y) - I\|$ would follow from this and the previous bounds on $\|\tau_{\exp_x(t\theta)\to x}E_\alpha(t,\theta) - te_\alpha\|$. Taking derivatives yield the bounds on $\|\partial_\alpha g_{\beta\gamma}\|$ and $\|\partial_{\alpha\beta}g_{\gamma\xi}\|$ in a similar but more complicated way. $\square$

Lemma 6. *Fix $x \in \mathcal{M}$ and let $\rho > 0$ be the injectivity radius at x. Define*

$$J(x,u) \;:=\; \big|\det \mathrm{d}\exp_x(u)\big| \;=\; \sqrt{\det g_{ij}(\exp_x u)}.$$

There exist universal constants $c, C > 0$, such that for $u \in T_x\mathcal{M}$ with $\|u\| \le \frac{c}{d}\min(\rho, 1/K)$, we have the following bound on $J(x,u)$:

$$\Big|J(x,u) - 1\Big| \le CdK\|u\|^2. \tag{6}$$

In particular, we have

$$\frac{1}{2} \le J(x,u) \le 2, \quad \|u\| \le \frac{c}{d}\min(\rho, 1/K).$$

Proof. Work in normal coordinates at x so that $\exp_x : B_{\mathrm{euc}}(0,\rho) \subset T_x\mathcal{M} \to B_{\mathrm{geo}}(x,\rho)$ is a diffeomorphism and $g_{ij}(0) = \delta_{ij}$, $\Gamma^k_{ij}(0) = 0$.

From Lemma 5, we know that $\|g(\exp_x u) - I\| \le CK\|u\|^2$. In the region $\|u\| \le \frac{c}{d}\min(\rho, 1/K)$, we have

$$\|g(\exp_x u) - I\| \le \frac{c}{d}.$$

Therefore, by Taylor expansion of determinants, we know

$$\begin{aligned}|\det g(\exp_x u) - 1| &= |\det(I + g(\exp_x u) - I) - 1| \\ &\le C\,\mathrm{tr}(g(\exp_x u) - I) \\ &\le Cd\|g(\exp_x u) - I\| \\ &\le Cd \cdot CK\|u\|^2.\end{aligned}$$

This concludes the proof by adjusting C if necessary. $\square$

We spell out the constants in a few classical inequalities in geometric analysis.

Lemma 7 (Schoen et al. (1994, Thm. 4.6))**.** *Let $(\mathcal{M}, g)$ be a complete Riemannian manifold with* $\mathrm{Ric}(\mathcal{M}) \ge -K$ *for some $K \ge 0$. Let $H(x,y,t)$ be the heat kernel, i.e., the fundamental solution of $(\Delta - \frac{\partial}{\partial t})u(x,t) = 0$. Then, for every $\delta_{\mathrm{Sch}} > 0$ and $\alpha > 1$,*

$$H(t,x,y) \;\le\; C(\delta_{\mathrm{Sch}}, d, \alpha)\, V_x\big(\sqrt{t}\big)^{-1/2} V_y\big(\sqrt{t}\big)^{-1/2}\, \exp\left[-\frac{r^2(x,y)}{(4+\delta_{\mathrm{Sch}})t} + C_1\,\delta_{\mathrm{Sch}}Kt\right],$$

where $V_x(R) = \mu(B_x(R))$, $C(\delta_{\mathrm{Sch}}, d, \alpha) = (1+\delta_{\mathrm{Sch}})^{d\alpha}\exp\big(\frac{1+\alpha}{\delta_{\mathrm{Sch}}}\big)$, and $C_1 = \frac{\alpha d}{\alpha - 1}$.

To unleash the power of Lemma 7, we need the following lower bound on volume of geodesic balls assuming bounded geometry.

Lemma 8 (Günther's comparison theorem)**.** *Under Assumption 1, we have*

$$V_x(r) \geq \frac{(2\pi)^{d/2}}{\Gamma(d/2)} \int_0^r \left(\frac{\sin(t\sqrt{K})}{\sqrt{K}} \right)^{d-1} \mathrm{d}t, \quad 0 \leq r \leq \rho.$$

Here Γ is the Gamma function. In particular, when $r \leq c \min(\rho, 1/\sqrt{K})$ for some small universal constant $c > 0$, we have

$$V_x(r) \geq \frac{\pi^{d/2}}{d\Gamma(d/2)} r^d \geq \frac{1}{d^{d/2}} r^d.$$

Proof. We observe that $\|\mathrm{Rm}\| \leq K$ implies that the sectional curvature is upper bounded by K, since by definition $\sec(u, v) = \mathrm{Rm}(u, v, u, v)$. The first inequality follows from the classical form of Günther's comparison theorem, see for example Gray (2003, Theorem 3.17). The second inequality follows from the elementary bound that $\sin(x) \geq \frac{1}{2}x$ for $x \in [0, c]$, where c is a small universal constant, and the crude bound $\Gamma(x) \leq x^{x-1}$ for $x \geq 1$. $\square$

A complementary lower bound to Lemma 7 is as follows.

Lemma 9 (Li & Xu (2011, Thm. 1.5))**.** *Let $(\mathcal{M}, g)$ be complete, possibly with $\mathrm{Ric}(\mathcal{M}) \geq -K$. For the (Neumann) heat kernel $H(x, y, t)$ and all $x, y \in \mathcal{M}$, $t > 0$,*

$$\begin{aligned} H(t, x, y) &\geq (4\pi t)^{-d/2} \frac{(2kt)^{d/2}}{(e^{2Kt} - 2Kt - 1)^{d/4}} \exp\left[-\frac{d(x, y)^2}{4t} \left(1 + \frac{Kt\coth(Kt) - 1}{Kt} \right) \right], \\ H(t, x, y) &\geq (4\pi t)^{-d/2} \exp\left[-\frac{d(x, y)^2}{4t} \left(1 + \tfrac{1}{3}Kt \right) - \frac{d}{4}Kt \right]. \end{aligned} \tag{7}$$

The above bounds for heat kernel translates seamlessly to p_t, since p_t is a convolution of p_0 with $H(t, \cdot, \cdot)$. We formalize this in the following lemma.

Lemma 10. *We have*

$$\inf_{x,y\in\mathcal{M}} H(t, x, y) \leq \inf_{x\in\mathcal{M}} p_t(x) \leq \sup_{x\in\mathcal{M}} p_t(x) \leq \sup_{x,y\in\mathcal{M}} H(t, x, y).$$

Proof. This follows from taking infimum and supremum respectively in the formula

$$p_t(y) = \int_{x\in\mathcal{M}} p_0(x) H(t, x, y) \mu(\mathrm{d}x).$$

$\square$

The following lemma compiles a few follow-ups of Li-Yau estimates (Hamilton, 1993; Han & Zhang, 2016) with constants made explicit.

Lemma 11. *Under Assumption 1, we have Han-Zhang's inequality*

$$\frac{\nabla^2 p_t}{p_t} \preceq C_{\mathsf{HZ}} \left(\frac{1}{t} + K \right) \left(1 + \log \frac{\sup p_{t/2}}{p_t} \right).$$

On the other hand, we also have Hamilton's Harnack inequality

$$\nabla^2 \log p_t = \frac{\nabla^2 p_t}{p_t} - \frac{(\nabla p_t)(\nabla p_t)^\top}{p_t^2} \succeq -\frac{1}{2t} g - C_{\mathsf{Ham}} \left(1 + \log \frac{\sup p_{t/2}}{p_t} \right) g$$

and

$$\|\nabla \log p_t\|^2 = \frac{\|\nabla p_t\|^2}{p_t^2} \leq C \left(\frac{1}{t} + K \right) \log \frac{\sup p_{t/2}}{p_t}.$$

Here $C > 0$ is a universal constant, $C_{\mathsf{HZ}} = CdK$, $C_{\mathsf{Ham}} = CdK^2$.

Proof. The last inequality follows from Hamilton (1993, Theorem 1.1).

The proof of the rest two inequalities require tracing the proofs of Hamilton (1993); Han & Zhang (2016). The details would be too tedious to reproduce here, so we leave pointers to relevant proofs for interested readers.

To prove the second inequality, we trace the proof of Hamilton (1993, Theorem 4.3) to see that if $A > 0$ is such that

$$\frac{\Delta_{\mathcal{M}} p_t}{p_t} \leq \frac{A}{t}\left(d + \log \frac{\sup p_{t/2}}{p_t}\right)$$

and

$$\frac{\|\nabla p_t\|^2}{p_t^2} \leq \frac{A}{t}\left(d + \log \frac{\sup p_{t/2}}{p_t}\right),$$

then

$$\frac{\nabla^2 p_t}{p_t} - \frac{(\nabla p_t)(\nabla p_t)^\top}{p_t^2} \succeq -\frac{1}{2t} g - CA(\|\mathrm{Rm}\| + \|\nabla \mathrm{Rm}\|) g.$$

Here $C > 0$ is an absolute constant. By Assumption 1, this implies

$$C_{\mathsf{Ham}} \leq CKC_{\mathsf{HZ}}.$$

Tracing the proof the main theorem in Han & Zhang (2016, Page 9), we can see

$$C_{\mathsf{HZ}} \leq C_{\mathsf{LY}}(1 + K) = CdK,$$

where C_{LY} is the maximum of the coefficients before t^{-1} and K in Li & Yau (1986, Theorem 1.2). This was explicitly defined as Cd therein, by setting $\alpha = 2$ there. This completes the proof. □

Lemma 12. *For any three points $x, y, z \in \mathcal{M}$, we have*

$$\frac{d(x,z)^2}{1-t} + \frac{d(z,y)^2}{t} \geq d(x,y)^2, \qquad \forall t \in (0,1).$$

Moreover, if x, y, z are within $\iota \leq 1/\operatorname{poly}(d, K, \rho^{-1})$ distance to each other, then the function

$$\psi(z) := \frac{d(x,z)^2}{1-t} + \frac{d(z,y)^2}{t} - d(x,y)^2$$

is $\frac{1-Cd^2K^2\iota}{t(1-t)}$-strongly convex in the normal coordinates at x (or y, z), where $C > 0$ is a universal constant.

Proof. The first inequality follows from Cauchy-Schwarz inequality and triangle inequality:

$$(1 - t + t)\left(\frac{d(x,z)^2}{1-t} + \frac{d(z,y)^2}{t}\right) \geq (d(x,z) + d(z,y))^2 \geq d(x,y)^2.$$

Proving the strong convexity requires Lemma 5 and Lemma 6, which implies that $d(x,z)^2$ and $d(z,y)^2$ are both $(1 - Cd^2K^2\iota)$-strongly convex in the normal coordinates. The desired conclusion then follows from

$$\frac{1}{1-t} + \frac{1}{t} = \frac{1}{t(1-t)}.$$

The proof is completed. □

B Initialization Error

We require the following result from Urakawa (2006, Proposition 2.6).

Lemma 13. *Denote $H(t,x,y)$ the heat kernel on $\mathcal{M}$. Assume*

$$A := \sup_{t \leq 1} \sup_{x \in \mathcal{M}} t^{d/2} H(t,x,x).$$

Then for any probability distribution p_0, its evolution along heat flow $\partial_t p_t = \frac{1}{2}\Delta_{\mathcal{M}}$ satisfies

$$\mathsf{TV}(p_t, \mu) \leq \sqrt{A}\, \mathrm{e}^{-\frac{1}{2\lambda_1}(t - \frac{1}{2})}, \quad t \geq 1.$$

We combine this with the Li-Yau upper bound (Lemma 7) to obtain

Lemma 14 (First part of Lemma 1)**.** *Under Assumption 1, there exists a universal constant $C > 0$ such that*

$$\mathsf{TV}(p_N, \mu) \le \mathrm{e}^{C(K+d\log(d/\rho))}\mathrm{e}^{-\frac{1}{2\lambda_1}(T-\frac{1}{2})}.$$

Proof. Plug the bound in Lemma 8 into Lemma 7 and use the fact that $\sup_{x,y} H(t,x,y)$ is decreasing in t (by convolution inequality), we obtain

$$\begin{aligned} A &\le (Cd)^d \min(\rho, 1/\sqrt{K})^d \mathrm{e}^{CK} \\ &\le \exp(C'K + C'd\log(d/\rho)), \end{aligned}$$

for some universal constants $C, C' > 0$. Note that we absorbed $d\log K$ into $K + d\log d$. The desired claim follows. □

C Score matching error

We now prove the second inequality in Lemma 1.

Lemma 15. *Under the same assumptions as in Theorem 1, we have*

$$\sum_{k=1}^{N}\int_{t_{k-1}}^{t_k} \mathbb{E}\|\mathscr{S}_{t_k,Y_{t_k}}(Y_t) - \mathscr{S}^\star_{t_k,Y_{t_k}}(Y_t)\|^2 \mathrm{d}t \le 2\varepsilon^2_{\mathsf{score}}.$$

Proof. This is relatively straightforward. Notice that in normal coordinates, by Lemma 5, we have

$$\begin{aligned} \|\mathscr{S}_{t_k,Y_{t_k}}(Y_t) - \mathscr{S}^\star_{t_k,Y_{t_k}}(Y_t)\|^2 &= g_{\alpha\beta}(Y_t)(s_\theta - \nabla\log p_t)^\alpha (s_\theta(t_k, Y_{t_k}) - \nabla \log p_{t_k}(Y_{t_k}))^\beta \\ &\le \|g(Y_t)\| \cdot \|s_\theta(t_k, Y_{t_k}) - \nabla\log p_{t_k}(Y_{t_k})\|^2 \\ &\le 2\|s_\theta(t_k, Y_{t_k}) - \nabla\log p_{t_k}(Y_{t_k})\|^2 \end{aligned}$$

for $d(Y_t, Y_{t_k}) \le \min(\frac{\rho}{3}, \frac{c}{K})$, and is 0 otherwise due to our cutoff η_ω. Therefore

$$\mathbb{E}\|\mathscr{S}_{t_k,Y_{t_k}}(Y_t) - \mathscr{S}^\star_{t_k,Y_{t_k}}(Y_t)\|^2 \le 2\mathbb{E}\|s_\theta(t_k, Y_{t_k}) - \nabla\log p_{t_k}(Y_{t_k})\|^2,$$

and consequently,

$$\begin{aligned} &\sum_{k=1}^{N}\int_{t_{k-1}}^{t_k} \mathbb{E}\|\mathscr{S}_{t_k,Y_{t_k}}(Y_t) - \mathscr{S}^\star_{t_k,Y_{t_k}}(Y_t)\|^2 \mathrm{d}t \\ &\le 2\sum_{k=1}^{N}(t_k - t_{k-1})\mathbb{E}\|s_\theta(t_k, Y_{t_k}) - \nabla\log p_{t_k}(Y_{t_k})\|^2 = 2\varepsilon^2_{\mathsf{score}}. \end{aligned}$$

This proves the claim, as desired. □

D Discretization error

Lemma 16. *Under the same assumptions as in Theorem 1 and assuming (3) without loss of generality, there is a universal constant $C > 0$ such that for $t_k - h \le t \le t_k$, we have*

$$\mathbb{E}\|\nabla\log p_t(Y_t) - \mathscr{S}^\star_{t_k,Y_{t_k}}(Y_t)\|^2 \le \frac{Cd^6K^8}{t^3}(t_k - t).$$

Proof. For convenience, set the reverse time

$$\tau := t_k - t.$$

The main challenge is that Li-Yau estimates provide sharp uniform control up to second-order derivatives of $\log p_t$, but a naïve calculation of the difference $\nabla\log p_t(Y_t) - \mathscr{S}^\star_{t_k,Y_{t_k}}(Y_t)$ involves third-order derivatives. More precisely, a straightforward Taylor expansion will introduce a factor of $\partial_\tau\nabla\log p_t$, which, by reverse-time heat equation $\partial_\tau p_t = -\frac{1}{2}\Delta_{\mathcal{M}} p_t$, contains third-order derivatives of p_t. We bypass this difficulty by making use of Itô's calculus to show that third-order derivatives cancel out; this is inspired by Benton et al. (2024), where a similar strategy was employed to the Euclidean setting.

Step I. Applying Itô/Stratonovish formula. We first compute $\partial_\tau \nabla \log p_t$. Since the forward heat equation is $\partial_t p_t = \frac{1}{2}\Delta_{\mathcal{M}} p_t$, we have

$$\partial_\tau \nabla \log p_t \;=\; -\partial_t \nabla \log p_t = -\nabla\left(\frac{\partial_t p_t}{p_t}\right) = -\frac{1}{2}\nabla\left(\frac{\Delta_{\mathcal{M}} p_t}{p_t}\right).$$

Use the manifold quotient rule and the identity $\Delta_{\mathcal{M}} \log p_t = \frac{\Delta_{\mathcal{M}} p_t}{p_t} - \|\nabla \log p_t\|^2$ to rewrite

$$\frac{\Delta_{\mathcal{M}} p_t}{p_t} = \Delta_{\mathcal{M}} \log p_t + \|\nabla \log p_t\|^2.$$

Therefore, we have

$$\begin{aligned}\partial_\tau \nabla \log p_t &= -\frac{1}{2}\nabla\left(\frac{\Delta_{\mathcal{M}} p_t}{p_t}\right) = -\frac{1}{2}\nabla \Delta_{\mathcal{M}} \log p_t - \frac{1}{2}\nabla\|\nabla \log p_t\|^2 \\ &= -\frac{1}{2}\nabla \Delta_{\mathcal{M}} \log p_t - \nabla^2 \log p_t \cdot \nabla \log p_t.\end{aligned}$$

On the other hand, it is straightforward to calculate

$$\nabla^2 \log p_t = \frac{\nabla^2 p_t}{p_t} - \frac{(\nabla p_t)(\nabla p_t)^\top}{p_t^2}.$$

Now, from Itô's formula, we know

$$\begin{aligned}&\mathrm{d}\nabla \log p_t(Y_t) \\ &= (\partial_\tau \nabla \log p_t)(Y_t)\mathrm{d}\tau + \nabla^2 \log p_t(Y_t)(\nabla \log p_t(Y_t)\mathrm{d}\tau + U_{Y_t} \circ \mathrm{d}W_t) + \frac{1}{2}\Delta_{\mathcal{M}} \nabla \log p_t(Y_t)\mathrm{d}\tau \\ &= -\frac{1}{2}\nabla \Delta_{\mathcal{M}} \log p_t \mathrm{d}\tau - \nabla^2 \log p_t \cdot \nabla \log p_t \mathrm{d}\tau + \nabla^2 \log p_t \cdot \nabla \log p_t \mathrm{d}\tau \\ &\quad + \nabla^2 \log p_t \cdot U_{Y_t} \circ \mathrm{d}W_t + \frac{1}{2}\Delta_{\mathcal{M}} \nabla \log p_t \mathrm{d}\tau \\ &= \frac{1}{2}(\Delta_{\mathcal{M}}\nabla - \nabla\Delta_{\mathcal{M}}) \log p_t \mathrm{d}\tau + \nabla^2 \log p_t \cdot U_{Y_t} \circ \mathrm{d}W_t \\ &= \frac{1}{2}\,\mathrm{Ric}^\sharp(\nabla \log p_t, \cdot)\mathrm{d}\tau + \nabla^2 \log p_t \cdot U_{Y_t} \circ \mathrm{d}W_t,\end{aligned}$$

where the last line follows from Bochner's identity $(\Delta_{\mathcal{M}}\nabla - \nabla\Delta_{\mathcal{M}})f = \mathrm{Ric}^\sharp(\nabla f, \cdot)$, and $\mathrm{Ric}^\sharp$ denotes the $(1,1)$-tensor obtained by raising one index in Ricci curvature. Notice here the cancellation of third-order derivatives.

On the other hand,

$$\mathrm{d}\mathscr{S}^\star_{t_k, Y_{t_k}}(Y_t) = \nabla \mathscr{S}^\star_{t_k, Y_{t_k}}(Y_t) \cdot (\nabla \log p_t \mathrm{d}\tau + U_{Y_t} \circ \mathrm{d}W_t) + \frac{1}{2}\Delta_{\mathcal{M}} \mathscr{S}^\star_{t_k, Y_{t_k}}(Y_t)\mathrm{d}\tau.$$

In normal coordinates, $\mathscr{S}^\star_{t_k, Y_{t_k}}$ is a constant vector field inside $B(0, \omega/3)$, therefore we have (cf. Lemma 4):

$$\nabla_\alpha \mathscr{S}^\star_{t_k, Y_{t_k}}(Y_t)^\beta = \Gamma^\beta_{\alpha\gamma} \mathscr{S}^\star_{t_k, Y_{t_k}}(Y_t)^\gamma = \Gamma^\beta_{\alpha\gamma} \nabla^\gamma \log p_{t_k}(Y_{t_k}), \quad d(Y_t, Y_{t_k}) \leq \omega/3,$$

and similarly, when $d(Y_t, Y_{t_k}) \leq \omega/3$, we have

$$\Delta_{\mathcal{M}} \mathscr{S}^\star_{t_k, Y_{t_k}}(Y_t)^\alpha = -\mathrm{Ric}^\alpha{}_\beta \nabla^\beta \log p_{t_k}(Y_{t_k}) - g^{\beta\gamma}\left(\partial_\gamma \Gamma^\alpha_{\beta\xi} + \Gamma^\alpha_{\gamma\zeta}\Gamma^\zeta_{\beta\xi} + \Gamma^\alpha_{\zeta\xi}\Gamma^\zeta_{\beta\gamma}\right)\nabla^\xi \log p_{t_k}(Y_{t_k}).$$

Step II. Bounding the coefficients. Combine the above formulas with the estimates given in Lemma 5, we obtain for some universal constant $C > 0$:

$$\begin{aligned}\left\|\mathrm{Ric}^\sharp(\nabla \log p_t, \cdot)\right\| &\leq CK\|\nabla \log p_t(Y_t)\|, \\ \left\|\nabla \mathscr{S}^\star_{t_k, Y_{t_k}}\right\| &\leq CK\|\nabla \log p_{t_k}(Y_{t_k})\|, \qquad d(Y_t, Y_{t_k}) \leq \omega/3,\end{aligned}$$

$$\left\|\Delta_{\mathcal{M}}\mathscr{S}^{\star}_{t_k,Y_{t_k}}\right\| \leq CK^2\|\nabla \log p_{t_k}(Y_{t_k})\|, \qquad\qquad d(Y_t, Y_{t_k}) \leq \omega/3.$$

Outside the geodesic ball $B_{Y_{t_k}}(\omega/3)$, the field $\mathscr{S}^{\star}_{t_k,Y_{t_k}}$ only inside $B_{Y_{t_k}}(\omega)$. Between these two balls, we have to take into account the radial derivative of the cutoff function η_ω, whose first order derivative is bounded by $C\omega^{-1}$ and second order derivative by $C\omega^{-2}$ by our construction (recall that $|\eta'| + |\eta''| \leq 100$). Apply Lemma 4 and Lemma 5 again, this time we bound

$$\begin{aligned}\left\|\nabla\mathscr{S}^{\star}_{t_k,Y_{t_k}}\right\| &\leq CK\omega^{-1}\|\nabla \log p_{t_k}(Y_{t_k})\|,\\ \left\|\Delta_{\mathcal{M}}\mathscr{S}^{\star}_{t_k,Y_{t_k}}\right\| &\leq CK^2\omega^{-2}\|\nabla \log p_{t_k}(Y_{t_k})\|.\end{aligned}$$

We now apply Itô's formula on manifold (i.e., take expectation and invoke the martingale property in (1)) and collect the above bounds to obtain (cf. Benton et al. (2024))

$$\begin{aligned}&\left|\frac{\mathrm{d}}{\mathrm{d}\tau}\mathbb{E}\|\nabla \log p_t(Y_t) - \mathscr{S}^{\star}_{t_k,Y_{t_k}}(Y_t)\|^2\right|\\
&\leq CK^2\left(\mathbb{E}\|\nabla \log p_t(Y_t)\|^2 + \mathbb{E}\|\nabla \log p_{t_k}(Y_{t_k})\|^2 + \mathbb{E}\|\nabla^2 \log p_t(Y_t)\|^2\right)\\
&\quad + CK^2\omega^{-2}\mathbb{E}\left[(\|\nabla \log p_t(Y_t)\|^2 + \|\nabla \log p_{t_k}(Y_{t_k})\|^2 + \|\nabla^2 \log p_t(Y_t)\|^2)\mathbf{1}_{d(Y_t,Y_{t_k})>\omega/3}\right]\\
&\leq CK^2\left(\mathbb{E}\|\nabla \log p_t(Y_t)\|^2 + \mathbb{E}\|\nabla \log p_{t_k}(Y_{t_k})\|^2 + \mathbb{E}\|\nabla^2 \log p_t(Y_t)\|^2\right)\\
&\quad + CK^2\omega^{-2}\sqrt{\mathbb{E}\|\nabla \log p_t(Y_t)\|^4 + \mathbb{E}\|\nabla \log p_{t_k}(Y_{t_k})\|^4 + \mathbb{E}\|\nabla^2 \log p_t(Y_t)\|^4}\sqrt{\mathbb{P}(d(Y_t,Y_{t_k}) > \omega/3)}\\
&\leq CK^2 d\left(\frac{1}{t} + dK^2\right)^2 \sup_{t\leq s\leq t_k}\sqrt{\mathbb{E}\log^4\frac{\sup p_{s/2}}{p_s(Y_s)}}\cdot\left(1 + \omega^{-2}\sqrt{\mathbb{P}(d(Y_t,Y_{t_k}) > \omega/3)}\right). \qquad (8)\end{aligned}$$

Here the last line follows from Lemma 11.

Step III. Controlling expectations via Chebyshev and Li-Yau estimates. To bound $\mathbb{E}\log^4\frac{\sup p_{s/2}}{p_s(Y_s)}$, we note that

$$\mathbb{E}\left(\frac{1}{p_t(Y_t)}\right) = \int \frac{1}{p_t}p_t\mathrm{d}\mu = \int \mathrm{d}\mu = 1.$$

By Chebyshev's inequality, we have

$$\mathbb{P}\left(\frac{1}{p_t(Y_t)} \geq \lambda\right) \leq \lambda^{-1}, \quad \lambda > 0,$$

and then

$$\mathbb{P}\left(\log^4\frac{1}{p_t(Y_t)} \geq \lambda\right) \leq \mathrm{e}^{-\sqrt[4]{\lambda}}, \quad \lambda \geq 0.$$

Integrate with respect to λ, we see

$$\mathbb{E}\log^4\frac{1}{p_t(Y_t)} \leq C.$$

We then apply Li-Yau's estimate (Lemma 7) combined with Lemma 8, Lemma 10 to obtain $\sup\log p_{s/2} \leq \sup\log H(s/2,x,y) \lesssim d\log\frac{d}{s} + Ks$, where the first inequality follows from $p_{s/2}$ being the convolution of p_0 with $H(s/2,x,y)$. These together shows

$$\mathbb{E}\log^4\frac{\sup p_{s/2}}{p_s(Y_s)} \leq \left(Cd\log\frac{d}{s} + CKs\right)^4 \leq \frac{Cd^5K^4}{\delta}.$$

Here we used $\log\frac{d}{s} \leq C\left(\frac{d}{s}\right)^{1/4}$, and $s \geq t \geq \delta$.

Step IV. Controlling exit probability via stopping time. It remains to bound the probability $\mathbb{P}(d(Y_t, Y_{t_k}) > \omega/3)$. This would follow from a stopping time argument. We claim that given $t_k - t \le h$, we have

$$\mathbb{P}(d(Y_t, Y_{t_k}) > \omega/3) \le \exp\left(-\frac{c\omega^2}{t_k - t}\right) \le \omega^4. \tag{9}$$

where the last inequality follows from (3). Plug this back into the desired conclusion of the lemma is proved.

We now prove (9). Let σ be the largest $t \le t_k$ such that $d(Y_t, Y_{t_k}) > \omega/3$. We have

$$\mathbb{P}(d(Y_t, Y_{t_k}) > \omega/3) \le \mathbb{P}(\sigma \ge t).$$

In the interval $[\sigma, t_k]$, Y_t stays in the geodesic ball $B_{Y_{t_k}}(\omega/3)$, and follows the SDE (2). In normal coordinates, this can be spelled out explicitly:

$$\begin{aligned} \mathrm{d}Y_t^\alpha &= \nabla^\alpha \log p_t(Y_t)\mathrm{d}t + A^\alpha_\beta(Y_t) \circ \mathrm{d}W_t^\beta \\ &= \left(\nabla^\alpha \log p_t(Y_t) + \frac{1}{2}(\partial_\gamma A^\alpha_\beta)A^{\beta\gamma}\right)\mathrm{d}t + A^\alpha_\beta \mathrm{d}W_t^\beta, \end{aligned} \tag{10}$$

where A is the square root of the matrix representing the coefficients of the Laplace-Beltrami operator

$$\frac{1}{\sqrt{\det g}}\partial_\alpha(\sqrt{\det g}\cdot g^{\alpha\beta}\partial_\beta).$$

It can be checked with the help of Lemma 5 that $\|A - I\| \le CdK\omega^2$, and $\|\partial_\alpha A\| \le Cd^2K\omega$; we omit the computation that has a similar pattern as many of the previous arguments. Furthermore, we have $\|\nabla \log p_t(Y_t)\| \lesssim (\delta^{-1} + K)\log \frac{\sup_{t/2}}{p_t}$ by the same argument via Lemma 11 as before. This time we combine the uniform bound provided by Lemma 9 with Lemma 7 to conclude $\log \frac{\sup p_{t/2}}{p_t} \lesssim (\delta^{-1} + K + d\log d)^2 \operatorname{Diam}(\mathcal{M})^2 \lesssim (\delta^{-1} + K + d\log d)^2K^2$ by Assumption 1.

Therefore, in view of Lemma 5 to convert the above bound to normal coordinates, and together with the aforementioned bound for A and $\partial_\alpha A$, we see that the drift term up to time σ will not exceed

$$\sup \|\nabla^\alpha \log p_t(Y_t) + \frac{1}{2}(\partial_\gamma A^\alpha_\beta)A^{\beta\gamma}\| \cdot (t_k - \sigma) \le C(\delta^{-1} + K + d\log d)^2K^2(t_k - \sigma) \le \frac{\omega}{12}, \tag{11}$$

where the last inequality used (3). On the other hand, the bound on A implies that the quadratic variation of the martingale part does not exceed

$$\int_\sigma^{t_k} A^\alpha_\gamma A^\gamma_\beta \mathrm{d}t \preceq 2(t_k - \sigma)I.$$

By Burkholder-Davis-Gundy inequality (Revuz & Yor, 2013), the tail of $\int_\sigma^{t_k} A^\alpha_\beta \mathrm{d}W_t^\beta$ is $O(1)$-subgaussian, thus we have

$$\mathbb{P}\left(\sigma \ge t, \left\|\int_\sigma^{t_k} A^\alpha_\beta \mathrm{d}W_t^\beta\right\| > \frac{\omega}{12}\right) \le \exp\left(\frac{-c(\omega - 2\sqrt{d(t_k - t)})^2}{t_k - t}\right).$$

In view of (3), combine this with (11) and (10), we have proved (9) as claimed. □

Lemma 17. *Under the same assumptions as in Theorem 1 and assuming (3) without loss of generality, the discretization error obeys the following upper bound:*

$$\sum_{k=1}^N \int_{t_{k-1}}^{t_k} \mathbb{E}\|\nabla \log p_t(Y_t) - \mathscr{S}^\star_{t_k, Y_{t_k}}(Y_t)\|^2 \mathrm{d}t \le \frac{Cd^6K^8}{\delta^3}h^2N,$$

where $C > 0$ is a universal constant.

Proof. This follows directly from Lemma 16. □

E BROWNIAN MOTION SIMULATION ERROR

In this section, we handle the Brownian motion simulation error using the machinery of Minakshisundaram-Pleijel parametrix. A complete introduction to this heavy machinery would require establish a whole system of notations and lemmas in geometric analysis, which is unduly burdensome. We instead refer the interested reader to Berline et al. (2003) for a comprehensive treatment, and point to results there whenever needed.

E.1 OVERVIEW

Our aim is to prove the following lemma.

Lemma 18. *Under the same assumptions as in Theorem 1, and assuming (3) without loss of generality, we have*

$$\mathsf{TV}(p_0^{\mathsf{aux}}, q_0^\star) \le \sqrt{hT}\ \mathrm{poly}(d, K, \rho^{-1}, \delta^{-1}).$$

To better explain the idea of the proof, we ignore the rejection sampling procedure in the construction of $\widehat{\mathsf{K}}_k$ temporarily. Our starting point is the observation that by Fokker-Planck equation, $\widehat{\mathsf{K}}_k$ is the heat kernel associated to the Euclidean Laplacian with drift $\mathscr{S}_{t_k,Y_{t_k}}$, in normal coordinates. On the other hand, $\mathsf{K}_k^{\mathsf{aux}}$ is also a heat kernel with the same drift, but associated to the manifold Laplace-Beltrami operator. The following lemma shows that the two solutions coincide up to first order in time, at least in a polynomially small neighborhood of initial point and in a polynomially short time.

Lemma 19. *Let $F^{\mathscr{S}}(t,x,y)$ be the (generalized) heat kernel for the operator $-\frac{1}{2}\Delta_{\mathcal{M}} + \langle \mathscr{S}_{t_k,Y_{t_k}}, \nabla\rangle$. Define the Euclidean density*

$$\varphi_t(u;x) \coloneqq \frac{1}{(2\pi t)^{d/2}} \exp\left(-\frac{\|u - \mathscr{S}_{t_k,Y_{t_k}}(x)t\|^2}{2t}\right), \quad u \in T_x\mathcal{M},$$

and let $\Phi(t,x,y)$ be the push forward by $\exp_x$ of the $\eta_\omega \varphi_t(\cdot;x)$, where η_ω is the cutoff function defined in (4). Then there exists some $\iota \le 1/\mathrm{poly}(d,K,\rho^{-1}) \le \omega$, such that for all

$$0 < t < \frac{\iota}{\mathrm{poly}(d,K,\rho^{-1})(1 + \|s_\theta(t_k, Y_{t_k})\|^2)},$$

we have

$$\left|\frac{F^{\mathscr{S}}(t, Y_{t_k}, y)}{\Phi(t, Y_{t_k}, y)} - 1\right| \le \mathrm{poly}(d,K,\rho^{-1})t.$$

With Lemma 19 in hand, it is tempting to calculate the KL error

$$\begin{aligned}
\mathsf{KL}(p_k^{\mathsf{aux}}\mathsf{K}_k^{\mathsf{aux}} \,\|\, p_k^{\mathsf{aux}}\widehat{\mathsf{K}}_k) &\lesssim -\mathbb{E}\int F^{\mathscr{S}}(h, Y_{t_k}, \cdot) \log \frac{\Phi(h, Y_{t_k}, \cdot)}{F^{\mathscr{S}}(h, Y_{t_k}, \cdot)} \\
&\lesssim \mathbb{E}\int F^{\mathscr{S}}(h, Y_{t_k}, \cdot)\left(\frac{F^{\mathscr{S}}(h, Y_{t_k}, \cdot)}{\Phi(h, Y_{t_k}, \cdot)} - 1\right)^2 \\
&\lesssim \mathrm{poly}(d,K,\rho^{-1})h^2,
\end{aligned}$$

where we ignore the fact that Lemma 19 holds only in a small neighborhood; the first line is post-processing inequality, and the second line follows from the fact that for two distributions p, q, we have

$$\begin{aligned}
\int p \log \frac{q}{p} &= \int p \log\left(1 + \frac{q-p}{p}\right) \\
&\ge \int p \left(\frac{q-p}{p} - C\frac{(q-p)^2}{p^2}\right) \\
&= -C \int p \left(\frac{q}{p} - 1\right)^2,
\end{aligned}$$

given $\frac{q}{p} - 1$ is sufficiently small, where the last line follows from $\int p = \int q = 1$. From this, we conclude that the accumulated error along N steps is bounded by $\mathrm{poly}(d,K,\rho^{-1})h^2 N = \mathrm{poly}(d,K,\rho^{-1})hT$, and the desired bound follows from Pinsker's inequality.

Apart from Lemma 19, the above computation is the essence of this proof. The rest of this section is devoted to formalizing this computation, and in particular, in handling exceptional events of exiting the polynomially small neighborhood.

E.2 A PARAMETRIX ESTIMATE

We present the proof of Lemma 19. For simplicity, denote by v^α the normal coordinate representation of $s_\theta(t_k, Y_{t_k})$. Naturally, our initial test solution is the drifted heat kernel, as simulated by our discretized process:

$$\varphi_t(u) \coloneqq \frac{1}{(2\pi t)^{d/2}} \exp\left(-\frac{\|u - vt\|^2}{2t}\right), \quad u \in \mathbb{R}^d.$$

Before we compare this with the manifold heat kernel, there is one subtlety we need to keep in mind. The density φ_t is with respect to the *Lebesgue* measure on $T_x\mathcal{M}$, not with respect to the *volume* on $\mathcal{M}$. In Berline et al. (2003), this is reflected by the distinction between the spaces $\Gamma(\mathcal{M}, \mathscr{E})$ and $\Gamma(\mathcal{M}, \mathscr{E} \otimes |\Lambda|^{1/2})$. We define and compute the corresponding density on $\mathcal{M}$ as follows:

$$\Phi(t, x, y) = \varphi_t(\log_x y)\sqrt{\Delta(x, y)}, \quad \Delta(x, y) \coloneqq \frac{|\det \mathrm{d} \log_x y|^2}{\det g(y)}, \quad y \in B_x(\rho).$$

Here all quantities are computed in normal coordinates. The factor $\Delta(x, y)$ is known as the van Vleck-Morette determinant. From Lemma 5 and Lemma 6, we know that

$$\frac{1}{2} \le \Delta(x, y) \le 2, \quad \text{if } d(x, y) \le \frac{c}{d}\min(\rho, 1/K). \tag{12}$$

We consider the generalized Laplacian

$$\mathcal{H} \coloneqq \frac{1}{2}\Delta_{\mathcal{M}} + \langle \mathscr{S}_{t_k, Y_{t_k}}, \nabla \rangle.$$

As in Lemma 19, denote by $F^{\mathscr{S}}$ the heat kernel of $\mathcal{H}$ at time $t_k - t_{k-1}$. The essence of the proof is a quantitative version of the following asymptotic expansion:

$$F^{\mathscr{S}}(t, x, y) = \Phi(t, x, y)(y) \cdot \left(1 + \sum_{i \ge 1} t^i u_i(x, y)\right), \quad \theta \to 0^+,$$

where u_i are smooth functions that can be computed explicitly via a recursive formula. We will not need this formula here, but instead require u_1 and the remainder terms to be bounded properly. Such bounds have been well-established, which we wrap up into the following lemma.

Lemma 20 (adapted from Berline et al. (2003))**.** *There exists a smooth function $u_1(x, y)$ on $\mathcal{M} \times \mathcal{M}$ such that*

$$\|u_1\|_{L^\infty(\mu)} + \|\nabla_y u_1\|_{L^\infty(\mu)} \le \mathrm{poly}(d, K, \rho^{-1})\|\mathscr{S}_{t_k, Y_{t_k}}\|_{L^\infty(\mathcal{M})},$$

and for all $0 < t \le 1/\mathrm{poly}(d, K, \rho^{-1})$, $y \in B_x(\rho/3)$, we have

$$\left|(\partial_t - H)\left[\Phi(t, x, y)(1 + t u_1(x, y))\right]\right| \le \mathrm{poly}(d, K, \rho^{-1})\|\mathscr{S}_{t_k, Y_{t_k}}\|^2_{L^\infty(\mathcal{M})}\Phi(t, x, y). \tag{13}$$

Proof. The first inequality follows from Theorem 2.26 in Berline et al. (2003), with H the same as our H and therefore $F = \langle \mathscr{S}_{t_k, Y_{t_k}}, \nabla \rangle$; all these coefficients are bounded by $\mathrm{poly}(d, K, \rho^{-1})\|\mathscr{S}_{t_k, Y_{t_k}}\|_{L^\infty(\mathcal{M})}$. The second inequality follows from the proof of Theorem 2.29, item (iii) in Berline et al. (2003). In particular, the role of the cutoff function ψ there is played by our η_ω, and with the notations there, we have

$$\left|(\partial_t - H)\left[\Phi(t, x, y)(1 + t u_1(x, y))\right]\right| \le \underbrace{\mathrm{poly}(\omega^{-1})\,\Phi(t, x, y)}_{\partial_y \eta_\omega \text{ related terms}} + t \cdot \Phi(t, x, y) \cdot |(\mathscr{B}_y u_2)(x, y)|.$$

Here $\mathscr{B}_y$ can be viewed as a coordinate-transformed version of H, applied to the variable y (precise definition can be found in the reference), and u_2 is the second order term in the expansion. Again, using Theorem 2.26 there, we can prove

$$\|\mathscr{B}_y u_2\|_{L^\infty(\mu)} \le \mathrm{poly}(d, K, \rho^{-1})\|\mathscr{S}_{t_k, Y_{t_k}}\|^2_{L^\infty(\mathcal{M})}.$$

The claimed bound follows from combining the above inequalities. □

Lemma 21 (Volterra series, Theorem 2.23 in Berline et al. (2003))**.** *Let*

$$\Phi_1(t,x,y) \coloneqq \Phi(t,x,y)(1+tu_1(x,y)),$$
$$r_1(t,x,y) \coloneqq (\partial_t - H)\,\Phi_1(t,x,y).$$

Define the time-space convolution operator

$$f * g(t,x,y) = \int_0^t \int f(t-s,x,z)g(s,z,y)\mu(\mathrm{d}z)\mathrm{d}s.$$

Then we have

$$F^{\mathscr{S}} = H_1 - H_1 * r_1 + H_1 * r_1 * r_1 - H_1 * r_1 * r_1 * r_1 \pm \cdots,$$

on any domain such that the series in the right hand side converges uniformly.

Lemma 22 (Iterative bounds for Volterra series)**.** *There exists a radius $\iota = 1/\operatorname{poly}(d,K,\rho^{-1}) \le \omega$ with universally constant coefficients such that the following holds. Assume $d(x,y) \le \iota$ and*

$$0 < t \le \frac{1}{\operatorname{poly}(d,K,\rho^{-1})(1+\|\mathscr{S}_{t_k,Y_{t_k}}\|^2_{L^\infty(\mathcal{M})})}.$$

Then we have

$$|H_1 * r_1| + |H_1 * r_1 * r_1| + |H_1 * r_1 * r_1 * r_1| + \cdots \le \operatorname{poly}(d,K,\rho^{-1})\|\mathscr{S}_{t_k,x}\|^2_{L^\infty(\mathcal{M})} t \cdot \Phi.$$

Proof. As in Berline et al. (2003), we first apply a cutoff function to localize to a normal neighborhood at x of radius $\iota \le \omega$. One may check that the bounds in Lemma 20 still holds, as the only change involved is replacing ρ with $\operatorname{poly}(d,K,\rho^{-1})$.

We then estimate

$$\begin{aligned}
&|H_1 * r_1(t,x,y)| \\
&= \left| \int_0^t \int_{B_x(\iota)} H_t(t-s,x,z) r_1(t,z,y)\mu(\mathrm{d}z)\mathrm{d}s \right| \\
&\le \operatorname{poly}(d,K,\rho^{-1})\|\mathscr{S}_{t_k,x}\|^2_{L^\infty(\mathcal{M})} \int_0^t \int_{B_x(\iota)} \Phi(t-s,x,z)\Phi(s,z,y)\mu(\mathrm{d}z)\mathrm{d}s \\
&\le \operatorname{poly}(d,K,\rho^{-1})\|\mathscr{S}_{t_k,x}\|^2_{L^\infty(\mathcal{M})} \int_0^t \int_{B_x(\iota)} \frac{1}{(2\pi(t-s))^{d/2}} \frac{1}{(2\pi s)^{d/2}} \mathrm{e}^{-\frac{d(x,z)^2}{2(t-s)} - \frac{d(z,y)^2}{2s}} \mu(\mathrm{d}z)\mathrm{d}s,
\end{aligned} \tag{14}$$

where the second line follows from Lemma 20, and the last line follows from the definition of Φ and the bound (12).

To proceed, we make use of Lemma 12. Let $z_\star$ be a minimizer of

$$\psi(z) = \frac{d(x,z)^2}{1-st^{-1}} + \frac{d(z,y)^2}{st^{-1}} - d(x,y)^2.$$

By obvious coercivity, when $t \le c\iota^2$ for some small universal constant $c > 0$, we have $z_\star \in B_x(4\iota)$. By Lemma 12 and Lemma 5, we have

$$\psi(z_\star) \ge 0, \quad \psi(z) \ge (1 - Cd^2K^2\iota)\frac{t^2}{s(t-s)} d(z,z_\star)^2.$$

Here the second inequality follows from strong convexity given by Lemma 12 and a comparison of geometric distance and Euclidean distance in normal coordinates fueled by Lemma 5. Denote for the moment that

$$\theta \coloneqq 1 - Cd^2K^2\iota.$$

Plug this back into (14), we obtain

$$|H_1 * r_1(t,x,y)|$$

$$
\begin{aligned}
&\leq \mathrm{poly}(d,K,\rho^{-1})\|\mathscr{S}_{t_k,x}\|^2_{L^\infty(\mathcal{M})}\int_0^t\int_{B_x(\iota)}\frac{1}{(2\pi(t-s))^{d/2}}\frac{1}{(2\pi s)^{d/2}}\mathrm{e}^{-\frac{\theta t d(z,z_\star)^2}{2s(t-s)}-\frac{d(x,y)^2}{2t}}\mu(\mathrm{d}z)\mathrm{d}s\\
&\leq \mathrm{poly}(d,K,\rho^{-1})\|\mathscr{S}_{t_k,x}\|^2_{L^\infty(\mathcal{M})}\mathrm{e}^{-\frac{d(x,y)^2}{2t}}\int_0^t\int_{B(0,\iota)}\frac{1}{(2\pi(t-s))^{d/2}}\frac{1}{(2\pi s)^{d/2}}\mathrm{e}^{-\frac{\theta t\|Z\|^2}{2s(t-s)}}\mathrm{d}Z\mathrm{d}s\\
&\leq \mathrm{poly}(d,K,\rho^{-1})\|\mathscr{S}_{t_k,x}\|^2_{L^\infty(\mathcal{M})}\frac{1}{(2\pi t)^{d/2}}\mathrm{e}^{-\frac{d(x,y)^2}{2t}}\theta^{-d/2}\int_0^t\mathrm{d}s\\
&\leq \mathrm{poly}(d,K,\rho^{-1})\|\mathscr{S}_{t_k,x}\|^2_{L^\infty(\mathcal{M})}\theta^{-d/2}\Phi(t,x,y)\cdot t,
\end{aligned}
$$

where the third line follows from change of variable to normal coordinates at $z_\star$ and use (12) to bound the determinant, the last line uses again the definition of Φ and (12), and the penultimate line follows from Gaussian integration

$$\int\frac{(\theta t)^{d/2}}{(2\pi s(t-s))^{d/2}}\mathrm{e}^{-\frac{\theta t\|Z\|^2}{2s(t-s)}}\mathrm{d}Z=1.$$

Now, we note that

$$\theta^{-d/2}=(1-Cd^2K^2\iota)^{-d/2}\leq\exp(2Cd^3K^2\iota)\leq 2,$$

give $\iota\leq\frac{1}{100Cd^3K^2}$. Putting these pieces together, we proved

$$|H_1*r_1|\leq t\cdot\underbrace{\mathrm{poly}(d,K,\rho^{-1})\|\mathscr{S}_{t_k,x}\|^2_{L^\infty(\mathcal{M})}}_{=:\kappa}\Phi.$$

Iterate this above process to see that the n-th convolution is bounded by

$$|H_1*r_1^{(j)}|\leq(\kappa t)^j\Phi.$$

Therefore,

$$\left|\sum_{j=1}^\infty(-1)^{j-1}H_1*r_1^{(j)}\right|\leq\sum_{j=1}^\infty(\kappa t)^j\Phi\leq\frac{\kappa t}{1-\kappa t}\Phi\leq 2t\Phi,$$

given $t\leq 1/(4\kappa)$. Recalling the definition of κ, the lemma is proved. □

We now have all the ingredients to prove Lemma 19.

Proof of Lemma 19. This follows immediately from Lemma 21 and Lemma 22. □

E.3 Handling exceptional events

Proof of Lemma 18. Denote

$$u(t,x,y)=\frac{1}{t}\left(\frac{\Phi(t,x,y)}{F^{\mathscr{S}}(t,x,y)}-1\right),\quad d(x,y)\leq\iota.$$

Then Lemma 19 implies

$$|u|\leq\mathrm{poly}(d,K,\rho^{-1}),\quad d(x,y)<\iota,\ \ 0<t<\frac{\iota}{\mathrm{poly}(d,K,\rho^{-1})(1+\|s_\theta(t_k,Y_{t_k})\|^2)}.$$

As a first simplification, we invoke (A3) in Assumption 1 to see

$$\|s_\theta(t_k,Y_{t_k})\|\leq\mathrm{poly}(d,K,\rho^{-1},\delta^{-1})(1+\|\nabla\log p_{t_k}(Y_{t_k})\|).$$

Then we can control $\|\nabla\log p_{t_k}(Y_{t_k})\|$ as in the proof of (11), resulting in

$$\|s_\theta(t_k,Y_{t_k})\|\leq\mathrm{poly}(d,K,\rho^{-1},\delta^{-1}).$$

Therefore, the conclusion of Lemma 19 simplifies to

$$|u|\leq\mathrm{poly}(d,K,\rho^{-1}),\quad d(x,y)<\iota,\ \ 0<t<\frac{\iota}{\mathrm{poly}(d,K,\rho^{-1},\delta^{-1})}.$$

Now, notice that the integration of $F^{\mathscr{S}}(h, Y_{t_k}, \cdot)$ outside $B_{Y_{t_k}}(\iota)$ is bounded by Lemma 7 and Lemma 8:

$$\int_{\mathcal{M}\setminus B_{Y_{t_k}}(\iota)} F^{\mathscr{S}}(h, Y_{t_k}, y)\mu(\mathrm{d}y) \le \exp\left(-\frac{c\iota^2}{h}\right), \tag{15}$$

and similarly

$$\int_{B_{Y_{t_k}}(\iota)} \Phi(h, Y_{t_k}, y)\mu(\mathrm{d}y) \ge 1 - \exp\left(-\frac{c\iota^2}{h}\right).$$

We may extend $u(h, \cdot, \cdot)$ to a bounded function with value in $[-\frac{1}{h}, h\exp(C\iota^2/h))$, defined on the whole $\mathcal{M}$ obeying the following two restrictions:

$$\int_{\mathcal{M}} u(h, Y_{t_k}, y)F^{\mathscr{S}}(h, Y_{t_k}, y)\mu(\mathrm{d}y) = 0, \tag{16}$$

$$\int_{\mathcal{M}\setminus B_{Y_{t_k}}(3\iota)} hu(h, Y_{t_k}, y)F^{\mathscr{S}}(h, Y_{t_k}, y)\mu(\mathrm{d}y) \le \exp\left(-\frac{c\iota^2}{h}\right). \tag{17}$$

The construction of such an extension is routine, for example, can be a constant function of order $h\exp(C\iota^2/h)$ outside $B_{Y_{t_k}}(2\iota)$, and tuned in $B_{Y_{t_k}}(2\iota) \setminus B_{Y_{t_k}}(\iota)$ to fulfill the restrictions.

With this extension, we find that

$$\widetilde{\mathsf{K}}_k := \widetilde{\Phi}(h, x, y) := (hu(h, x, y) + 1)F^{\mathscr{S}}(h, x, y)$$

is a kernel on $\mathcal{M}$, by (16). Moreover, (15) and (17) together imply

$$\mathsf{TV}(p_{Y_{t_k}}\widehat{\mathsf{K}}_k,\ p_{Y_{t_k}}\widetilde{\mathsf{K}}_k) \le \exp\left(-\frac{c\iota^2}{h}\right).$$

Therefore, by post-processing inequality,

$$\mathsf{TV}(p_N\widetilde{\mathsf{K}}_N\cdots\widetilde{\mathsf{K}}_1,\ q_0^\star) \le N\exp\left(-\frac{c\iota^2}{h}\right) \le N\exp(-c'h^{-1/2}) \le h, \tag{18}$$

by choosing $\iota = \frac{1}{10}h^{1/4}$ and $h \le \frac{1}{\mathrm{poly}(d,K,\rho^{-1},\delta^{-1})}$.

On the other hand, the KL computation following Lemma 19, together with the bound $u \le h\exp(C\iota^2/h)$ so that $\log\frac{\widetilde{\Phi}(h,x,y)}{F^{\mathscr{S}}(h,x,y)} \le C\iota^2/h$, show that

$$\begin{aligned}\mathsf{KL}(p_0^{\mathsf{aux}} \,\|\, p_N\widetilde{\mathsf{K}}_N\cdots\widetilde{\mathsf{K}}_1) &\le \sum_{k=1}^N \mathrm{poly}(d, K, \rho^{-1}, \delta^{-1})h^2 + N^2\frac{C\iota^2}{h}\exp\left(\frac{-c\iota^2}{h}\right)\\ &\le \mathrm{poly}(d, K, \rho^{-1}, \delta^{-1})h^2N\\ &\le \mathrm{poly}(d, K, \rho^{-1}, \delta^{-1})hT,\end{aligned}$$

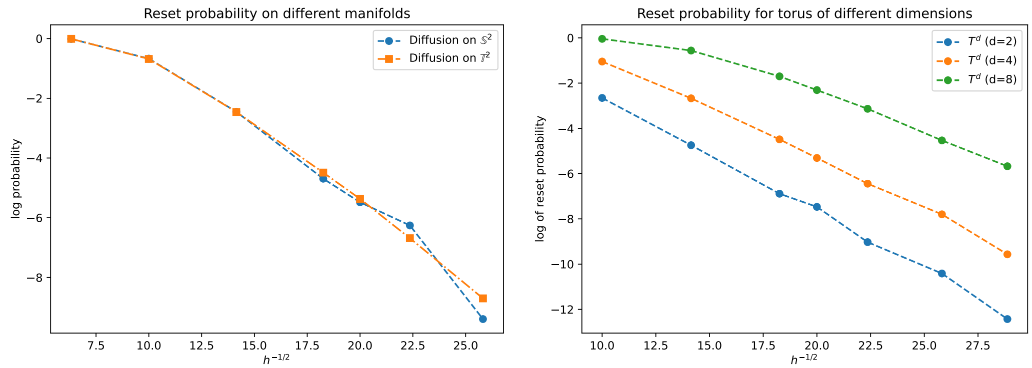
where the penultimate line follows from choosing $h \le \frac{\iota^2}{\mathrm{poly}(d,K,\rho^{-1},\delta^{-1})}$, and the last line follows from the choice of schedule $hN = T - \delta \le T$. The final TV bound follows from combining the above estimates and Pinsker's inequality. $\square$

F PROOF OF MAIN RESULTS

Proof of Lemma 1. This follows from combining Lemma 14 and Lemma 15. $\square$

Proof of Lemma 2. This follows from Lemma 17 and our choice of schedule $hN = T - \delta \le T$. $\square$

Proof of Theorem 1. This follows from Lemma 1, Lemma 2, and Lemma 18. $\square$



(a) (b)

Figure 1: Reset probabilities on spheres and tori. In Figure 1a, we examine the relationship between $h^{-1/2}$ and the log of the reset probability of Algorithm 1 on both sphere $\mathbb{S}^2$ and torus $\mathbb{T}^2$ under the reset rules of Algorithm 1. In both cases, we see that the reset probability decays exponentially, confirming the conclusion in (18). In Figure 1b, we examine the same statistics on high-dimensional tori, and we find increasing d only shifts the curves to the right but leaves the exponential decay rate in $h^{-1/2}$ unchanged.

G Simulations

In this section, we verify the results in Theorem 1 on compact manifolds by measuring the exit probability in the reverse steps of Algorithm 1 and the TV distance between the target distribution (a Gaussian mixture) and the recovered distribution.

Reset probability on $\mathbb{S}^2$ and $\mathbb{T}^2$. We start by examining the total reset probability on the unit 2-sphere $\mathbb{S}^2$ and on the 2-torus $\mathbb{T}^2$. We run the backward process in Algorithm 1 with different stepsizes h in each setting with p_0 being Gaussian mixture (see below for definition), and record the fraction of trials whose tangent update $\Delta_k = h\, s_{t_k}(Y_k) + \sqrt{h}\, G_k$ violates $\|\Delta_k\| \leq h^{1/4}$ at least once among all steps. On both manifolds, Figure 1a shows a clear linear trend of the logarithm of reset probability of against $h^{-1/2}$, consistent with our exponentially small tails.

High-dimensional torus. We extend this experiment to the d-dimensional flat torus $\mathbb{T}^d$ for different stepsizes. Figure 1b reports the logarithm of the reset probability versus $h^{-1/2}$ for $d \in \{2, 4, 8\}$. Increasing d raises the baseline reset rate, yet the slope of the decay remains essentially unchanged—resets remain exponentially rare as $h \downarrow 0$, which aligns with our analysis.

TV accuracy on $\mathbb{T}^d$ with warped Gaussian mixture. Finally, we assess distributional accuracy in TV for a *warped Gaussian mixture* target on $\mathbb{T}^d$, $d \in \{1, 2, 3\}$. Here the warped Gaussian distribution is defined to be the push-forward of the Gaussian distribution under the universal covering $\mathbb{R}^d \to \mathbb{T}^d$ given by $(x_1, \cdots, x_d) \mapsto (\mathrm{e}^{i2\pi x_1}, \cdots, \mathrm{e}^{i2\pi x_d})$, and warped Gaussian mixture is similarly the push-forward of Gaussian mixture in $\mathbb{R}^d$. The result is depicted in Figure 2, which confirms that the total variation decays fast with the increase of the number of steps.

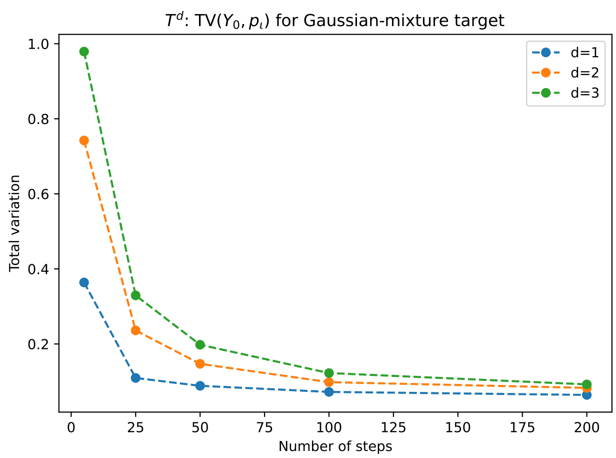


Figure 2: TV distance on $\mathbb{T}^d$ with a warped Gaussian-mixture target. The total variation is estimated with a kernel density estimator.