Total Variation Rates for Riemannian Flow Matching

Yunrui Guan
Department of Applied Mathematics and Statistics
Johns Hopkins University
yguan34@jh.edu

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

Shiqian Ma
Data Science and AI Institute,
and Department of Applied Mathematics and Statistics,
Johns Hopkins University
sqma@jhu.edu

Abstract

Riemannian flow matching (RFM) extends flow-based generative modeling to data supported on manifolds by learning a time-dependent tangent vector field whose flow-ODE transports a simple base distribution to the data law. We develop a nonasymptotic Total Variation (TV) convergence analysis for RFM samplers that use a learned vector field together with Euler discretization on manifolds. Our key technical ingredient is a differential inequality governing the evolution of TV between two manifold ODE flows, which expresses the time-derivative of TV through the divergence of the vector-field mismatch and the score of the reference flow; controlling these terms requires establishing new bounds that explicitly account for parallel transport and curvature. Under smoothness assumptions on the population and learned flow-matching field, together with mean-square approximation guarantees for the learned field, we obtain explicit bounds of the form $\mathrm{TV} \le C_{\mathrm{Lip}}\, h + C_\varepsilon\, \varepsilon$, cleanly separating numerical discretization and learning errors. Here, h is the step-size and ε is the target accuracy. Instantiations yield *explicit* polynomial iteration complexities on the hypersphere S^d, and on the Hadamard manifolds (for e.g., SPD manifolds) under mild moment conditions.

1 Introduction

Flow-based generative modeling offers a conceptually clean sampling paradigm by representing a complex target distribution as the endpoint of a continuous-time transport from a simple reference law. In *flow matching*, one fixes a family of interpolating measures $(\pi_t)_{t\in[0,1]}$ between an easy base distribution π_0 and the data distribution π_1, and learns a time-dependent vector field whose flow pushes π_0 to π_1. When the data lie on a Riemannian manifold (M, g), particles evolve according to an intrinsic ODE $\dot{X}_t = v_t(X_t)$, and the associated densities satisfy the manifold continuity equation $\partial_t p_t + \mathrm{div}_g(p_t v_t) = 0$, yielding a sampler that respects the geometry by construction. Crucially, because Riemannian Flow Matching (RFM) is purely deterministic and relies only on integrating intrinsic ODEs, it avoids discretizing manifold-valued Brownian motion as required by Riemannian diffusion models [10], where practical simulation is considerably more delicate.

Despite their ever-increasing practical usage [42, 35, 31, 32, 8, 29], a central theoretical question is unexplored: *how fast* does a discretized RFM sampler with a learned velocity field converge to the target distribution, in distributional metrics? Our main contribution is to provide the first *nonasymptotic Total Variation (TV) bounds* for RFM samplers on both compact and Hadamard manifolds.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

We develop intrinsic stability estimates for solutions of the Riemannian continuity equation under perturbations of the driving vector field, yielding explicit control of $\mathrm{TV}(\hat{\pi}_T, \pi_T)$ for $T = 1 - \delta$, ($\delta > 0$) in terms of (i) the approximation quality of the learned field $\hat{v}_t$ to an ideal field v_t in geometry-aware norms, (ii) regularity of the dynamics (e.g., bounds on covariant derivatives and divergence), and (iii) geometric characteristics of (M, g) that govern volume distortion. We then combine these stability estimates with a careful analysis of discretized flows to obtain end-to-end guarantees for practical samplers, isolating the statistical and computational contributions. The resulting theory clarifies which properties of the learned velocity field are truly required for strong sampling accuracy on manifolds, and how curvature and anisotropy enter the constants and rates, providing a principled foundation for understanding Riemannian flow matching algorithms.

Related works. Generative modeling on Riemannian manifolds was introduced by [10], who generalized score-based generative modeling [34] from Euclidean space to Riemannian manifolds, and proved an error bound of sampling error. The error bound in [10] fails to be polynomial, which is caused by technical difficulties in analyzing discretization error in manifold Brownian motion simulation. Other works including [15] and [28] also studied generative modeling on mainfolds without theoretical guarantees.

This difficulty is related to sampling on Riemannian manifolds via diffusion processes. Early work analyzed KL divergence along Langevin diffusions, e.g., on hyperspheres [24] and Hessian manifolds [11]. Later, [7] used a geometric approach with coupling to obtain W_1 bounds; see also [18] for Lie groups and [12] for high-accuracy proximal sampling. These results highlight that Riemannian Brownian motion—the basic component of Riemannian diffusion models—is challenging both theoretically and computationally: discretization analyses are largely geometric and typically give Wasserstein error bounds (e.g., [7]), while exact simulation is generally unavailable since its transition density lacks a closed form except in special cases. This in turn complicates denoising score matching and sample generation for Riemannian diffusion models. More recently, [40, Lemma 19] established a polynomial discretization error bound in total variation distance. Building on this, [40] further proved the first polynomial iteration complexity for Riemannian score-based generative models, but the established bounds are only qualitative and do not reveal the precise dependencies on the problem parameters.

Another line of research is the Riemannian Flow Matching method, first proposed in [5] and later explored by [6] and [39]. We remark that, due to its deterministic formulation, RFM does not require access to the heat kernel or simulation of manifold Brownian motion, thereby avoiding the main bottleneck of Riemannian diffusion models.

In the Euclidean setting, prior works established discretization error bounds for probability flow ODEs and flow matching under various metrics, including TV [25, 16], Wasserstein [3, 2, 44, 13], and KL divergence [36]. Without contraction or dissipativity assumptions, these bounds generally fail to achieve polynomial dependence on Lipschitz constants; fully polynomial rates are known mainly in TV distance [23, 25]. Related polynomial guarantees for stochastic interpolants were obtained in [27]. Recent works also studied low-dimensional and manifold statistical structure in Euclidean FM [33, 43]. To our knowledge, no prior work provides discretization error analysis for flow matching on general Riemannian manifolds.

2 Basics and problem formulation

2.1 Riemannian geometry basics

Let (M, g) be a simply connected d-dimensional Riemannian manifold, not necessarily realized as a submanifold of Euclidean space, where g denotes the Riemannian metric. Unless otherwise specified, $\|\cdot\|$ denotes the norm induced by g. We use T_xM to denote tangent space at location $x \in M$, and remark that $\|\cdot\|$ is well-defined on each tangent space T_xM. We let Exp to denote exponential map, and also define the inverse of the exponential map, i.e., the logarithm map: Given $X_0, X_1 \in M$, $\mathrm{Log}_{X_0}(X_1) \in T_{X_0}M$ is such that $\mathrm{Exp}_{X_0}(\mathrm{Log}_{X_0}(X_1)) = X_1$. We use P_x^y to denote the parallel transport from x to y, along the minimizing geodesic, unless otherwise specified. Based on this, we can define geodesic interpolation as follows.

$$X_t := \mathrm{Exp}_{X_0}(t\,\mathrm{Log}_{X_0}(X_1)) \tag{1}$$

Note that $X_t' = P_{X_0}^{X_t} \operatorname{Log}_{X_0}(X_1) = \frac{1}{1-t} \operatorname{Log}_{X_t}(X_1)$.

We use Grad to denote the Riemannian gradient, Div the Riemannian divergence, and ∇ the Levi–Civita connection on (M, g). For a smooth function f, the gradient Grad f is the unique vector field satisfying $\langle \text{Grad } f,\, v\rangle = df(v)$ for all $v \in T_xM$. For a smooth vector field u, the divergence is defined by Div $u := \operatorname{tr}(\nabla u)$; equivalently, for any local g-orthonormal frame $\{e_i\}_{i=1}^d$, Div $u(x) = \sum_{i=1}^d \langle \nabla_{e_i} u,\, e_i \rangle_x$. For a vector field u and a tangent vector $v \in T_xM$, the covariant derivative $\nabla_v u(x)$ is defined as follows: choose any smooth local vector field V with $V(x) = v$, and set $\nabla_v u(x) := (\nabla_V u)(x)$; this is well-defined (i.e., independent of the extension V) because ∇ is $C^\infty(M)$-linear in its first argument. The map $\nabla u(x)$ may be viewed as a $(1,1)$-tensor (an endomorphism) $\nabla u(x) : T_xM \to T_xM$ defined by $(\nabla u(x))(v) := \nabla_v u(x)$. Accordingly, we define the operator norm induced by g as $\|\nabla u(x)\|_{\text{op}} := \sup_{v \neq 0} \frac{\|\nabla_v u(x)\|}{\|v\|} = \sup_{\|v\|=1} \|\nabla_v u(x)\|$.

We now provide some intuition on Levi–Civita connection and covariant derivatives. On a manifold, the tangent spaces T_xM and T_yM at different points are distinct vector spaces, so one cannot subtract vectors at x and y directly. The Levi–Civita connection ∇ provides a canonical way to *differentiate* vector fields by comparing nearby tangent vectors in a manner that is compatible with the metric ($\nabla g = 0$) and has no torsion. Along a curve γ, the induced covariant derivative $D_t := \nabla_{\gamma'(t)}$ plays the role of a directional derivative: it measures the intrinsic rate of change of a vector field along γ after accounting for the variation of tangent spaces, and it is the notion of derivative that makes "constant velocity" along γ coincide with the geodesic equation.

Algorithm 1 RFM with Early Stopping

1: Denote the learned vector field by $\hat{v}(t, x)$.
2: Choose step size schedule $\{h_i\}$ and early stopping time $T \in [0, 1)$ such that $\sum_{i=0}^{N-1} h_i = T$, where N is the total number of steps.
3: Perform Euler discretization for N steps:
4: Generate $x_0 \sim \pi_0$, set $t_0 = 0$.
5: **for** $k = 0$ to $N - 1$ **do**
6: $\quad x_{t_{k+1}} = \operatorname{Exp}_{x_{t_k}}(h_k \hat{v}(t_k, x_{t_k}))$.
7: $\quad t_{k+1} \leftarrow t_k + h_k$.
8: **end for**
9: Output a sample $x_T := x_{t_N} \sim \hat{\pi}_T$.

Let $x \in M$ and $B_a(0)$ be the open metric ball in (T_xM, g_x). The injectivity radius at x, denoted $\operatorname{Inj}(x)$, is defined as the maximum radius a s.t. the exponential map $\operatorname{Exp}_x\big|_{B_a(0)} : B_a(0) \subset T_xM \to M$ is a diffeomorphism onto its image. The injectivity radius of M is $\operatorname{Inj}(M) := \inf_{x \in M} \operatorname{Inj}(x)$.

Let R denote the (Riemann) curvature tensor associated with ∇, defined for smooth vector fields X, Y, Z by $R(X,Y)Z := \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z$. For linearly independent $u, v \in T_xM$, the sectional curvature of the plane $\sigma = \operatorname{span}\{u, v\}$ is

$$K(\sigma) := \frac{\langle R(u,v)v,\, u\rangle_x}{\|u\|^2\|v\|^2 - \langle u, v\rangle_x^2}$$

and in particular, if $u \perp v$ and $\|u\| = \|v\| = 1$, then $K(u, v) = \langle R(u,v)v,\, u\rangle_x$.

Let $\gamma : [a, b] \to M$ be a geodesic, and write $D_t := \nabla_{\gamma'(t)}$ for the covariant derivative along γ. A vector field J along γ is called a *Jacobi field* if it satisfies the Jacobi equation $D_t^2 J + R(J, \gamma')\gamma' = 0$. Let $p = \gamma(a)$ and $q = \gamma(b)$. We say that p and q are *conjugate along* γ if there exists a nonzero Jacobi field J along γ such that $J(a) = J(b) = 0$.

For a unit-speed geodesic $\gamma : [0, \infty) \to M$ with $\gamma(0) = p$, define its *cut time* $t_{\text{cut}}(\gamma) := \sup\{t > 0 : \gamma|_{[0,t]}$ minimizes distance between its endpoints$\}$. If $t_{\text{cut}}(\gamma) < \infty$, the *cut point of p along γ* is $\gamma(t_{\text{cut}}(\gamma))$. The *cut locus* of p, denoted $\operatorname{Cut}(p)$, is the set of all cut points of p over all unit-speed geodesics emanating from p. It is a classical fact that along any geodesic starting at p, the cut time occurs no later than the first conjugate time (if a *conjugate point* occurs at all); moreover, a cut point may occur strictly before the first conjugate point when there are multiple minimizing geodesics to the same endpoint.

We now provide some intuition on the above concepts. The curvature tensor quantifies how covariant derivatives fail to commute; it encodes the intrinsic bending of the manifold. Sectional curvature reduces this information to a 2-dimensional direction σ: positive sectional curvature tends to make nearby geodesics in σ focus toward each other, while negative curvature tends to make them spread apart. Jacobi fields describe the infinitesimal separation between nearby geodesics (they arise from variations of geodesics), so zeros of a Jacobi field correspond to a loss of local uniqueness/minimality

of geodesics. Conjugate points describe a differential property, corresponding to the degeneracy of $d\exp$, hence naturally enter comparison theory used for our analysis later (see Section A.2). On the other hand, cut points describe a geometric property, describing when geodesics cease to be minimizing, as well as when Log_x stops being single-valued.

2.2 Riemanian flow matching

Let π_0 and π_1 be two probability distributions supported on a Riemannian manifold M. The Riemannian Flow Matching (RFM) framework aims to learn a time-dependent vector field $v(t, x)$ whose induced flow transports samples from π_0 to π_1. Concretely, RFM seeks v such that if $X_0 \sim \pi_0$, then the solution $\{X_t\}_{t\in[0,1]}$ to the ODE

$$dX_t = v(t, X_t)\, dt, \qquad X_0 \sim \pi_0, \tag{2}$$

satisfies $X_1 \sim \pi_1$. The ODE exhibits a maximal solution as long as v is continuous in time and locally-Lipschitz in space; see, for example [19, Chapter IV]. Moreover, if the manifold is complete and $v(t, x)$ satisfies a linear growth bound, the solution exists globally in time. We also remark here that [38] provided more refined conditions that hold more generally for the case of Euclidean spaces.

A key geometric constraint is that for every (t, x), the velocity must lie in the *tangent space at the current point*:

$$v(t, x) \in T_xM, \qquad t \in [0, 1].$$

That is, $v(t, x)$ represents a valid instantaneous direction of motion *along the manifold* at x. This is where the Riemannian setting differs fundamentally from the Euclidean one. When $M = \mathbb{R}^d$, all tangent spaces are canonically identical, i.e., $T_x\mathbb{R}^d \cong \mathbb{R}^d$ for all x, so we can view the vector field simply as a global map $v : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$ with a single, fixed vector space as codomain.

On a general manifold M, however, the tangent spaces $\{T_xM\}_{x\in M}$ form a family of *different vector spaces attached to different points*. Although each T_xM has the same dimension, there is no canonical identification between T_xM and T_yM when $x \neq y$. As a consequence, the "output space" of $v(t, \cdot)$ depends on x: v is a section of the tangent bundle rather than a map into a single fixed $\mathbb{R}^d$. This location-dependence forms a main source of technical differences between Euclidean flow matching and RFM, and it will matter in the analysis later.

In flow matching, one typically parameterizes the time-dependent vector field with a neural network and learns it by minimizing the *conditional flow matching* objective. In the Riemannian setting, choosing the coupling between the endpoints to be the *independent coupling*, i.e., sampling $X_0 \sim \pi_0$ and $X_1 \sim \pi_1$ independently, yields the training loss

$$\min_u \ \mathbb{E}_{t,\ X_0\sim\pi_0,\ X_1\sim\pi_1}\big[\|u(X_t, t) - P^{X_t}_{X_0}\,\mathrm{Log}_{X_0}(X_1)\|^2\big], \tag{3}$$

where $\mathrm{Log}_x(\cdot)$ is the Riemannian logarithm map, and P^b_a denotes parallel transport along the interpolation curve from a to b. In Euclidean space, the standard choice is the straight-line interpolation $X_t := tX_1 + (1-t)X_0$ between X_0 and X_1; on a Riemannian manifold, this naturally generalizes to the geodesic interpolation X_t in (1). We denote the population minimizer by $v(t, x)$ and the learned (neural) approximation by $\hat{v}(t, x)$.

It is useful to note that the population minimizer admits a closed-form expression as a conditional expectation (see, e.g., [13]). Specifically, for any fixed $t < 1$ and $x \in M$,

$$v(t, x) = \frac{1}{1-t}\,\mathbb{E}[\mathrm{Log}_x(X_1) \mid X_t = x] = \frac{1}{1-t}\int_M \mathrm{Log}_x(x_1)\, p_t(x_1 \mid x)\, dV_g(x_1), \tag{4}$$

where $p_t(x_1 \mid x)$ denotes the conditional density of X_1 given $X_t = x$ under the above independent coupling and the chosen interpolation, and dV_g is the Riemannian volume measure.

To generate samples, we numerically simulate the learned flow-matching ODE in Algorithm 1. While the ideal (population) flow driven by the true minimizer v would transport π_0 to π_1 at time $t = 1$, in practice we only have access to a learned approximation $\hat{v}$ and a time-discretized integrator. A key difficulty is the behavior of the conditional flow-matching vector field near the terminal time. Note from (4) that the population minimizer contains an explicit $1/(1-t)$ factor. Along the *exact* geodesic bridge used to define X_t, the term $\mathbb{E}[\mathrm{Log}_x(X_1) \mid X_t = x]$ typically scales like $(1-t)$ so that $v(t, x)$ remains finite; however, this cancellation is fragile. Any mismatch between the learned field $\hat{v}$ and

the population field v, or any deviation of the simulated trajectory from the training interpolation, can destroy the cancellation and make the effective dynamics increasingly stiff as $t \to 1$ (large velocities and/or large sensitivities with respect to x).

This stiffness directly motivates *early stopping*. In Algorithm 1, we use $\{h_i\}$ to denote the step size schedule. In the constant step size case, we simply have $h_i \equiv h, \forall i$. With a pre-designed step size schedule and early stopping time T, the algorithm generates $\{x_{t_k}\}$ where $t_k := \sum_{i=0}^{k-1} h_i$ and consequently by definition $t_N = T$. The Euler discretization accumulates both numerical discretization error and statistical/approximation error from learning $\hat{v}$. Near $t = 1$, the factor $\frac{1}{1-t}$ amplifies errors in estimating the small conditional mean $\mathbb{E}[\mathrm{Log}_x(X_1) \mid X_t = x]$: a small absolute error in this conditional expectation can translate into a much larger error in the velocity, which then produces an $O(1)$ perturbation over the last few integration steps and shifts the terminal law away from π_1. For this reason, following existing probability-flow analyses in the Euclidean setting (e.g., [25, 23]) we establish convergence rates to an *approximate* distribution at a terminal time $T < 1$, corresponding to stopping the integration before the most ill-conditioned regime.

2.3 Challenges in the Riemannian setting

We now summarize the main technical challenges to establish our results, that are absent in the Euclidean setting.

- **Well-definedness of the interpolated ODE.** In the discretization analysis, for each $t \in [t_k, t_{k+1})$, let y_t denote the spatial location of Euler interpolation. We need to uniquely recover the footpoint y_k such that the Euler interpolation starting from y_k reaches y_t. In Euclidean space, this follows from a direct differential argument; see, for example, [27, Lemma B.1]. On a Riemannian manifold, however, this is no longer automatic. To write the Euler interpolation as an ODE, the frozen velocity has to be moved to the tangent space at the current point, and we must show that the current point can be traced back to a unique starting point. Proving this requires geometric tools including geodesic variations and Jacobi field estimates; see Lemmas 5 and 16.
- **Smoothness and the cut locus.** In Euclidean space, the smoothness of the conditional density $p_t(x_1 \mid x_t = x)$ is trivial. On compact manifolds, this is no longer automatic because the cut locus may introduce sharp corners when we apply the change of variable formula. For hyperspheres, we resolve this issue explicitly and verify the smoothness needed for the population vector field; see Lemma 20 and 22.
- **Regularity beyond model spaces.** On Euclidean spaces and hyperspheres, many derivative bounds can be obtained from explicit formulas; e.g., see Lemmas in Appendix D.1 for hyperspheres. In contrast, for general Hadamard manifolds and SPD manifolds, such formulas are typically unavailable. The analysis therefore requires intrinsic bounds for derivatives of the exponential and logarithm maps in curved settings (see e.g., Lemma 44 and Proposition 54), which form one of the main technical ingredients of our argument.

3 Main results

In this section, we establish error bounds in terms of total variation distance. First, we discuss our assumptions and present our error bound for general manifolds under our assumptions. Then, we specialize to two different cases: Hadamard manifolds that is also a symmetric space (which are non-compact) and hyperspheres (which are compact), where we obtain explicit iteration complexity results. We start with the following assumption on the curvature of the manifold.

Assumption 1 (Assumptions on Riemannian Manifold)**.** *M is a complete manifold without boundary. There exists $L_R \geq 0$ s.t. $\|R(u,v)v\| \leq L_R\|u\|\|v\|^2, \forall u, v \in T_xM$, and all sectional curvatures of M are bounded below by $K_{\min}$ and bounded above by $K_{\max}$.*

This assumption imposes standard global geometric regularity on M. Completeness ensures that geodesics and the exponential map are well defined for all times. The bound $\|R(u,v)v\| \leq L_R\|u\|\|v\|^2$ controls the curvature operator, and the two-sided sectional curvature bound $K_{\min} \leq K \leq K_{\max}$ rules out both excessive focusing and overly rapid divergence of geodesics. These properties are crucial for the stability and comparison estimates used later.

Assumption 2 (Regularity of the true Vector Field)**.** *Let $(X_t)_{t\in[0,1)}$ denote the population RFM trajectory satisfying $dX_t = v(t, X_t)\, dt$ with $X_0 \sim \pi_0$. The true vector field $v(t, x)$ is at least C^2 in t, x and satisfies $\forall t \in [0, 1)$:*

(1) $\mathbb{E}_{X_t}\|\nabla v(t, X_t)\|^2 \le (L_t^{v,x})^2$ *for* $\|\cdot\|_{\mathrm{op}}$ *and* $\|\cdot\|_{\mathrm{HS}}$*, and* $\mathbb{E}_{X_t}\left\|\frac{d}{dt}v(t, X_t)\right\|^2 \le (L_t^{v,t})^2$.
(2) $\mathbb{E}_{X_t}\left|\frac{d}{dt}\,\mathit{Div}\; v(t, X_t)\right| \le L_t^{\mathit{Div},t}$*, and* $\mathbb{E}_{X_t}\|\,\mathit{Grad}_x\,\mathit{Div}\; v(t, X_t)\|^2 \le (L_t^{\mathit{Div},x})^2$.
(3) $\mathbb{E}_{X_t}\|v(t, X_t)\|^2 \le (L_t^v)^2$.

The above conditions require the population vector field $v(t, x)$ to be smooth and sufficiently regular in expectation so that the induced flow depends stably on time and initial conditions. The bounds on $\|\nabla v(t, x)\|$, $\|v(t, x)\|$, and $\left\|\frac{d}{dt}v(t, x)\right\|$ control spatial Lipschitz regularity, magnitude, and time variation of the dynamics. In Euclidean space $M = \mathbb{R}^d$, $\nabla v(t, x)$ reduces to the Jacobian $D_x v(t, x)$, so $\|\nabla v(t, x)\|_{\mathrm{op}} = \|D_x v(t, x)\|_{2\to 2}$. The bounds on $\|\,\mathrm{Grad}_x\,\mathrm{Div}\; v(t, x)\|$ and $\left|\frac{d}{dt}\,\mathrm{Div}\; v(t, x)\right|$ control the spatial and temporal variation of Div $v(t, x)$, and are needed to bound the divergence term in the TV-based discretization error analysis.

Assumption 3 (Regularity of Learned Vector Field)**.** *Assume that the learned vector field $\hat{v}$ satisfies: (1)* $\|\hat{v}(t, x)\| \le L_t^{\hat{v}}$*, (2)* $\|\nabla\hat{v}(t, x)\| \le L_t^{\hat{v},x}$ *and (3)* $\|\,\mathit{Grad}\;\mathit{Div}\;\hat{v}(t, x)\| \le L_t^{\mathit{Div}\;\hat{v},x}$*, where $L_t^{\hat{v}}, L_t^{\hat{v},x}, L_t^{\mathit{Div}\;\hat{v},x}$ are of the same order as $L_t^v, L_t^{v,x}, L_t^{\mathit{Div},x}$.*

Here, Assumption 3 is of essential importance to guarantee the well-behavedness of the sampling ODE. The condition that the Lipschitz constants of the learned vector field are of the same order of the true field is made purely for convenience to avoid a more complicated looking bound.

Assumption 4 (Estimation Error)**.** *There exists some small $\varepsilon > 0$ s.t. the following hold $\forall t \in \{t_k\}_{k=0}^{N-1}$: (1)* $\mathbb{E}_{X_t}[\|\hat{v}(t_k, X_{t_k}) - v(t_k, X_{t_k})\|^2] \le \varepsilon^2$*, (2)* $\mathbb{E}_{X_t}[|\mathit{Div}\;\hat{v}(t_k, X_{t_k}) - \mathit{Div}\; v(t_k, X_{t_k})|] \le \varepsilon$.

Assumption 3 and 4 are in line with state-of-the-art TV-based discretization analyses for Euclidean ODE generative models, see, e.g., [27, 16]. In our setting, the first item controls vector-field estimation error, while the second is used in Lemma 12 to control the expected absolute difference between Div $\hat{v}(t_k, X_{t_k})$ and Div $v(t_k, X_{t_k})$, which contributes directly to the TV bound. With separate tolerances ε_v and $\varepsilon_{\mathrm{div}}$, the result retains the same structure but contains two corresponding error terms. We use $\varepsilon = \max\{\varepsilon_v, \varepsilon_{\mathrm{div}}\}$ to simplify notation. Recent work [41] on Jacobian estimation error also provides supporting evidence for the plausibility of such a divergence estimation assumption. Below, we present our rates.

Theorem 1 (Sampling Error)**.** *Let Assumptions 1, 2, 3 and 4, hold, and further assume Inj $(M) > 0$. Picking the step size h to satisfy the requirements in Lemma 5 and Lemma 18, we have*

$$\mathrm{TV}\,(\pi_T, \hat{\pi}_T) \le h\mathsf{C}_{\mathit{Lip}} + \varepsilon\mathsf{C}_{\mathit{eps}},$$

where $\mathsf{C}_{\mathit{Lip}}$ and $\mathsf{C}_{\mathit{eps}}$ defined in (10) *are functions of constants assumed in assumptions.*

We impose the step-size restriction in Lemma 5 to ensure that each Euler step stays in a regime where $F_{t_k,h}(x) = \mathrm{Exp}_x(h\hat{v}(t_k, x))$ remains invertible, so that the continuous-time interpolation and the associated vector field $\tilde{v}$ are well defined; see Section 4 for further discussion.

The constant $\mathsf{C}_{\mathsf{Lip}}$ collects the regularity moduli that govern how local Euler errors propagate through the flow, and hence controls the $\mathcal{O}(h)$ discretization term. In contrast, $\mathsf{C}_{\mathsf{eps}}$ collects the corresponding stability factors for the estimation error $\hat{v} - v$ and its divergence, quantifying how an ε estimation error propagates through the continuity equation.

Theorem 1 gives a general sampling error bound under abstract regularity assumptions. Propositions 2 and 3 instantiate this bound on Hadamard symmetric spaces and hyperspheres, respectively, by verifying Assumption 2 and making the resulting iteration complexity explicit.

Remark. *The early stopping error can be controlled in the Wasserstein distance (see Definition 8) as: $W_1(\pi_1, \pi_T) \lesssim 1 - T$. On a compact manifold, TV bound implies a W_1 bound, as $W_1(\pi_T, \hat{\pi}_T) \le \mathrm{diam}\,(M)\,\mathrm{TV}\,(\pi_T, \hat{\pi}_T)$ [37, Theorem 6.15]. Thus for compact manifolds, we can convert our TV bound into a bound in W_1 distance, and obtain an error bound for $W_1(\hat{\pi}_T, \pi_1)$, considering both sampling error and early stopping error.*

3.1 Iteration complexity for Hadamard manifolds that is also a symmetric space

We now instantiate Theorem 1 to Hadamard manifolds with $\nabla R = 0$ by verifying Assumption 2. We then specialize the result to $\mathrm{SPD}(n) := \{X \in \mathbb{R}^{n\times n} : X \succ 0, X^T = X\}$ equipped with the affine-invariant metric $g_X(U,V) = \mathrm{tr}(X^{-1}UX^{-1}V)$, a setting arising in many applications [26, 8]. Proposition 2 below gives the resulting iteration complexity. To obtain these rates, we compute the order of the regularity constants in Proposition 14 and analyze error propagation under a specific step-size schedule. Following [23, 25, 27], we focus on the well-estimated scenario $\varepsilon \approx 0$ and report the number of sampling steps N required to achieve $\varepsilon_{\text{target}}$ accuracy with respect to π_T.

Proposition 2. *Let M be a Hadamard manifold that is also a symmetric space. We adopt early stopping, i.e., simulate the ODE on the interval $[0,T] \subsetneq [0,1]$. We choose the prior distribution to be a Riemannian Gaussian distribution centered at some $z \in M$: $\pi_0(x) \propto \exp\big(-d\,d(x,z)^2\big)$. We further assume the data distribution π_1 satisfies*

$$\max\big\{\mathbb{E}[d(X_1,z)^2 e^{24\lambda d(X_1,z)}], \mathbb{E}[e^{24\lambda d(X_1,z)}]\big\} \le M_\lambda, \quad \textit{where} \quad \lambda = \max\{1,\kappa\}, \tag{5}$$

for some constant M_λ. Under Assumption 1, 3 and Assumption 4, with a carefully designed step size schedule as in (12)*, the iteration complexity to reach ε_{target} accuracy in TV distance is*

$$N = \mathcal{O}\Big(\frac{\sqrt{\frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}}}\, d^{4+12\lambda} L_R^3 M_\lambda}{\varepsilon_{target}} \log^2 \frac{1}{1-T}\Big).$$

In particular, when $M = \mathrm{SPD}(n)$ for which the dimension $d = n(n+1)/2$, choose prior distribution as $\pi_0(x) \propto \exp(-\frac{n(n+1)d_g(x,I)^2}{2})$ (here we set the center of π_0 as $I \in \mathrm{SPD}(n)$ for notation simplicity). Then the iteration complexity can be reduced to $N = \mathcal{O}(\frac{d^{16} L_R^3 M_\lambda}{\varepsilon_{target}} \log^2 \frac{1}{1-T})$.

To see how geometry enters this result, we consider the model function $s_{K_{\min}}(r) = \sinh(\kappa r)/\kappa$ which describes the growth of Jacobi fields in the constant sectional curvature $K_{\min} = -\kappa^2$ hyperbolic model. At a high level, when the manifold is not of constant curvature, we can estimate regularity constants by quantities involving $s_{K_{\min}}(r)$ through comparison theorems. Because $s_{K_{\min}}(r)$ grows exponentially in r and distances are unbounded on a Hadamard manifold, controlling these regularity bounds in expectation requires the exponential-moment condition (5) in Proposition 2. This condition plays a role analogous to finite-moment conditions (e.g., $\mathbb{E}[\|X_1\|^4] < \infty$ and stronger conditions like bounded support) appeared in Euclidean discretization analysis [27, 25, 44]. The symmetric-space condition plays a separate role: $\nabla R = 0$ eliminates the curvature-derivative terms in the higher-order variation equations, allowing these regularity estimates to be expressed in a relatively simple way.

We remark that although the order of d is large, we expect it could be improved further by more refined computation and additional assumptions. Part of this large exponent comes from how our proof bounds the exponential growth of $s_{K_{\min}}(r)$ with distance by powers of d. In model spaces such as hyperbolic space, $s_{K_{\min}}(r)$ arises from an explicit Jacobi-field calculation, so the exponential growth therefore reflects actual negative-curvature distortion rather than a proof artifact.

3.2 Hyperspheres

For compact manifolds, we provide explicit rates on the d-dimensional hypersphere as a concrete instantiation of Theorem 1: we verify Assumption 2 and convert the resulting bound into explicit iteration complexity. Geometry enters through the sine Jacobian and the cut-locus indicator in the conditional density (see Lemma 20). Then in Lemma 22 we show that the sine Jacobian factor vanishes at the support boundary, making its product with the indicator sufficiently smooth to establish the required C^2 regularity of $v(t,x)$. A detailed discussion on general compact manifolds is deferred to Section 4.2.1.

Proposition 3. *For some finite $d \ge 4$, let $M = \mathcal{S}^d$, and $T < 1$ be the early stopping time. Assume π_1 has a smooth positive density. Pick $\pi_0(x)$ to be the uniform distribution on the hypersphere. Under Assumption 1, 3 and Assumption 4, to reach ε_{target} accuracy in TV distance: (i) for constant step size as in* (11)*, the complexity is $N = \mathcal{O}(d^2/\varepsilon_{target}(1-T))$; (ii) for a carefully designed step size schedule as in* (12)*, it can be improved to $N = \mathcal{O}(\frac{(d-1)^2}{\varepsilon_{target}} \log^2 \frac{1}{1-T})$.*

For hyperspheres, Assumption 2 can in fact be verified pointwise, without taking expectation. As a result, Proposition 3 also holds under a uniform estimation error assumption, which is stronger than Assumption 4. The trade-off is that the resulting iteration complexity depends on the density ratio M_1/m_1, where $0 < m_1 \le \pi_1(x) \le M_1$ for all $x \in M$.

4 Proof sketch and intermediate results

We outline the proof of the main TV-rate bound. Both the population flow-matching dynamics and its numerical approximation are viewed as deterministic transports on the manifold driven by time-dependent vector fields, with randomness entering only through the initialization. Our argument has three steps:

1. (*TV stability under transport*) derive a differential identity for $\partial_t \,\mathrm{TV}\,(p_t, q_t)$ in terms of the mismatch between the driving vector fields (Lemma 4);
2. (*Continuous-time interpolation of Euler*) construct an interpolation of the Euler scheme that is itself an ODE on M, so that Lemma 4 applies;
3. (*Term-by-term control*) bound the resulting right-hand side using score regularity, vector-field regularity, and a curvature-dependent estimate for the divergence of the interpolated field.

Integrating the differential inequality over time then yields a bound on the terminal sampling error $\mathrm{TV}\,(X_1, Y_1)$ (or, when needed, the limit $t \uparrow 1$).

We begin with the key TV-derivative lemma, the Riemannian analogue of [25, Lemma 4.2]. Its proof proceeds by writing the evolution of p and q through the continuity equation, differentiating $\int (p-q)_+$, and integrating by parts. All operators are interpreted intrinsically in the Riemannian sense. The second line of (8) follows by taking absolute values and applying Cauchy–Schwarz to the inner-product term.

Lemma 4. *Let X_t, Y_t be stochastic processes on M, satisfying the following ODE:*

$$dX_t = v(t, X_t)\,dt, \qquad X_0 \sim p_0 \tag{6}$$

$$dY_t = \tilde{v}(t, Y_t)\,dt, \qquad Y_0 \sim q_0 \tag{7}$$

Let $p(t,x)$ be the law of X_t and $q(t,x)$ be the law of Y_t. Define $\Omega_t = \{x \in M : p(t,x) - q(t,x) > 0\}$. Then we have

$$\begin{aligned}\frac{\partial \,\mathrm{TV}\,(X_t, Y_t)}{\partial t} &= \int_{\Omega_t} p(t,x)\left(\mathit{Div}\,(\tilde{v}(x,t) - v(x,t)) + \langle \mathit{Grad}\,\log p(t,x), \tilde{v}(x,t) - v(x,t)\rangle\right) dV_g(x) \\ &\le \mathbb{E}[|\,\mathit{Div}\,(\tilde{v}(x,t) - v(x,t))|] + \mathbb{E}[\|\,\mathit{Grad}\,\log p(t,x)\|\,\|\tilde{v}(x,t) - v(x,t)\|]. \end{aligned} \tag{8}$$

4.1 Continuous time interpolation

We now explain how Lemma 4 is used to analyze the Euler discretization error. We take (6) to be the flow-matching ODE driven by the (population) vector field $v(t,\cdot)$, and we want (7) to represent the numerical method. A direct interpolation of the Euler scheme would be

$$y_t \;=\; \mathrm{Exp}_{y_k}\big((t - t_k)\hat{v}(t_k, y_k)\big), \qquad t \in [t_k, t_{k+1}),$$

that is, the velocity is frozen at the left endpoint. However, this is not immediately an ODE of the form $dy_t = \tilde{v}(t, y_t)\,dt$, since the frozen vector $\hat{v}(t_k, y_k)$ lies in $T_{y_k}M$ rather than $T_{y_t}M$. The natural remedy is to *transport the frozen velocity along the curve*. We define

$$\tilde{v}(t, y_t) := P_{y_k}^{y_t}\hat{v}(t_k, y_k) \in T_{y_t}M, \qquad t \in [t_k, t_{k+1}),$$

so that $dy_t = \tilde{v}(t, y_t)\,dt$. To apply Lemma 4, we need $\tilde{v}$ to be defined as a function of (t,x) rather than implicitly through a trajectory. This requires invertibility of the Euler map $F_{t,h}(x) := \mathrm{Exp}_x(h\hat{v}(t,x))$. In particular, for $t \in [t_k, t_{k+1})$, the Euler interpolation can be written as $F_{t_k, t-t_k}(x_k) = \mathrm{Exp}_{x_k}((t - t_k)\hat{v}(t_k, x_k))$. Lemma 5 shows that for sufficiently small h, the map $F_{t_k,h}$ is invertible. In that case, we can express the interpolation vector field at an arbitrary point x by pulling back to the unique preimage $x_k = F_{t_k,t-t_k}^{-1}(x)$ and then parallel-transporting the frozen velocity to T_xM:

$$\tilde{v}(t,x) = P^x_{F^{-1}_{t_k,t-t_k}(x)}\hat{v}(t_k, F^{-1}_{t_k,t-t_k}(x)),$$

and along a trajectory (x_t) of (7), we have $F^{-1}_{t_k,t-t_k}(x_t) = x_{t_k}$.

Lemma 5. *Let M be simply connected Riemannian manifold that satisfies Assumption 1. Let b be any vector field on M, satisfying $\|b(x)\| \le B, \forall x \in M$. Assume $\|\nabla b(x)\| \le L_\nabla$. Let $R = \mathit{Inj}\,(M)$. To guarantee $F_{t_k, t-t_k}$ being invertible, we require*

$$h < \min\{\frac{R}{B}, \frac{1}{4L_\nabla}, \sqrt{\frac{3}{4\|b(x)\|^2 L_R(2 + 2L_\nabla \max\{\frac{1}{\sqrt{K_{\min}}}, 1\})}}\}, \qquad \textit{if} \quad K_{\min} > 0,$$

$$h < \min\{\frac{R}{B}, \frac{1}{4L_\nabla}, \sqrt{\frac{3}{4\|b(x)\|^2 L_R(2\frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}} + 4\frac{\cosh(\sqrt{-K_{\min}})-1}{-K_{\min}} L_\nabla)}}\}, \quad \textit{if} \quad K_{\min} < 0,$$

$$h < \min\{\frac{R}{B}, \frac{1}{4L_\nabla}, \sqrt{\frac{3}{4\|b(x)\|^2 L_R(2 + hL_\nabla)}}\}, \qquad \textit{if} \quad K_{\min} = 0.$$

The invertibility requirement ensures that a single Euler step remains in a regime where the exponential map does not fold the manifold (e.g., by crossing conjugate points or exiting the injectivity radius), so that the map $x \mapsto \mathrm{Exp}_x((t - t_k)\hat{v}(t_k, x))$ stays injective. Otherwise, distinct preimages could be mapped to the same point. Lemma 5 prevents this by combining injectivity-radius control with bounds on $dF_{t,h}$ via curvature-related comparison and covariant-derivative estimates. When $\mathrm{Inj}\,(M) = \infty$ (for example, on $\mathbb{R}^d$ or $\mathrm{SPD}\,(n)$), the injectivity-radius constraint becomes vacuous and the condition simplifies accordingly.

4.2 Discretization analysis

Although (6) and (7) are deterministic ODEs, the random initialization makes (X_t) and (Y_t) stochastic processes through their induced laws. To isolate *discretization* error, we couple them by taking the *same* initial distribution: $X_0 \sim \pi_0$, and $Y_0 \sim \pi_0$. Then $\mathrm{TV}\,(X_t, Y_t)$ measures the discrepancy between the population flow-matching transport and the transport induced by the Euler interpolation. In particular, at terminal time (or in the limit $t \uparrow 1$), $\mathrm{TV}\,(X_1, Y_1)$ is exactly the sampling error attributable to time discretization (and, if present, to the use of $\hat{v}$ instead of v).

Applying Lemma 4 with $p(t, \cdot)$ the law of X_t and $q(t, \cdot)$ the law of Y_t, and integrating in time, yields

$$\mathrm{TV}\,(X_1, Y_1) \le \mathrm{TV}\,(X_0, Y_0) + \int_0^1 (\mathbb{E}[|\,\mathrm{Div}\,(\tilde{v}(t, X_t) - v(t, X_t))|] + \mathbb{E}[\|\,\mathrm{Grad}\,\log p_t(X_t)\|\|\tilde{v}(t, X_t) - v(t, X_t)\|])dt.$$

Since $\mathrm{TV}\,(X_0, Y_0) = 0$ under the shared initialization, the task reduces to controlling the two error terms on the RHS. We do so by bounding $\mathbb{E}\big[\|\,\mathrm{Grad}\,\log p_t(X_t)\|^2\big]$, $\mathbb{E}\big[\|\tilde{v}(t, X_t) - v(t, X_t)\|^2\big]$, and $\mathbb{E}\big[|\,\mathrm{Div}\,(\tilde{v}(t, X_t) - v(t, X_t))|\big]$. We apply Cauchy–Schwarz to the the second expectation in (8):

$$\mathbb{E}[\|\,\mathrm{Grad}\,\log p_t(X_t)\|\|\tilde{v}(t, X_t) - v(t, X_t)\|] \le \sqrt{\mathbb{E}[\|\,\mathrm{Grad}\,\log p_t(X_t)\|^2]}\sqrt{\mathbb{E}[\|\tilde{v}(t, X_t) - v(t, X_t)\|^2]}.$$

Thus, once we establish suitable bounds for these three terms (uniformly for t away from 1, and with explicit dependence on t as $t \uparrow 1$), we can integrate over time to obtain a quantitative TV-rate.

4.2.1 Score regularity and vector field regularity

We now summarize the regularity inputs needed to control the score term and the vector-field mismatch. The score bound is the main "density regularity" ingredient, while the vector-field bounds ensure that (i) the Euler interpolation is well defined and (ii) the mismatch $\tilde{v} - v$ can be quantified in geometry-aware norms.

- **Score regularity.** We show that there exists a (possibly time-dependent) constant L_t^{score} such that, for all $t \in [0, 1)$, $\mathbb{E}\big[\|\,\mathrm{Grad}\,\log p_t(X_t)\|^2\big] \le L_t^{\mathrm{score}}$. The bound is allowed to deteriorate as $t \uparrow 1$; this is unavoidable in general and matches the behavior of population conditional flow fields near the terminal time. We establish this in Proposition 6 for compact manifolds and in Proposition 7 for Hadamard manifolds (both stated below).
- **Regularity on compact manifolds.** On compact manifolds, curvature is bounded and distances are uniformly controlled, but the presence of cut points can obstruct smoothness of conditional

densities such as $p_t(x_1 \mid x_t = x)$, which enter the explicit formulae for the population flow field. We verify Assumption 2 on $\mathcal{S}^d$ in Appendix D by exploiting explicit computations on the sphere. For more general compact geometries, controlling smoothness across the cut locus is substantially more delicate and is the main reason we phrase the required assumptions abstractly.

- **Regularity on Hadamard manifolds.** On Hadamard manifolds, the absence of cut points guarantees global smoothness of the exponential map and of the relevant conditional densities, which greatly simplifies differentiability issues. Pointwise bounds for derivatives of $v(t, x)$ typically involve factors that are Poly (d), Poly $(1/(1-t))$ and Poly $(s_{K_{\min}}(d(x_0, x_1)))$. We verify Assumption 2 for Hadamard manifolds with $\nabla R \equiv 0$ in Appendix E by combining curvature comparison with the moment condition (5).

Proposition 6. *Let M be a compact manifold satisfying Assumption 1. Under Assumption 2, consider π_0 being uniform distribution, exists some finite number L_t^{score} that depends on t s.t.* $\mathbb{E}\big[\| \textit{Grad} \log p_t(X_t)\|^2\big] \le L_t^{\text{score}}, \forall t \in [0, 1)$.

Proposition 7. *Let M be a Hadamard manifold that is also a symmetric space, with sectional curvature satisfying $-\kappa^2 = K_{\min} \le 0$. For prior being chosen as $X_0 \sim e^{-dd(x_0,z)^2}$, we have* $\mathbb{E}\big[\| \textit{Grad} \log p_t(X_t)\|^2\big] \lesssim M_\lambda d^2 L_R^2 d^{12\lambda}/(1-t)^2$*, where λ and M_λ are as in* (5)*.*

4.2.2 The divergence term

Finally, we comment on the divergence contribution $\mathbb{E}\big[|\operatorname{Div}(\tilde{v}(t, X_t) - v(t, X_t))|\big]$. In Euclidean space, the Euler interpolation simply freezes the velocity and no parallel transport is required; consequently, Div $\tilde{v}$ can be handled with elementary calculus. On a manifold, by contrast, our interpolated field $\tilde{v}$ is defined through (i) *parallel transport* and (ii) an *implicit inverse mapping* $F^{-1}_{t_k, t-t_k}$. Both operations contribute nontrivially to $\nabla\tilde{v}$ and hence to Div $\tilde{v}$.

Our approach is to explicitly differentiate the representation $\tilde{v}(t, x) = P^x_{F^{-1}_{t_k,t-t_k}(x)}\, \hat{v}\big(t_k, F^{-1}_{t_k,t-t_k}(x)\big)$ for $t \in [t_k, t_{k+1})$, along the interpolation geodesic under a parallel orthonormal frame, and decompose Div $\tilde{v}(t, x) = \sum_{i=1}^d \big\langle \nabla_{E_i(t)}\tilde{v}(t, x),\, E_i(t)\big\rangle$ into two pieces: (i) derivatives of $\hat{v}(t_k, \cdot)$ evaluated at the preimage point, and (ii) curvature distortion from differentiating parallel transport. Combining these estimates and using the curvature bounds in Assumption 1, we obtain a bound on Div $\tilde{v}$ and hence on Div $(\tilde{v} - v)$, stated precisely in Lemma 18. This divergence control, together with the score/vector-field regularity results above, completes the term-by-term bounds needed to integrate (8) and prove the stated TV-rate.

5 Conclusion and limitations

We established total variation sampling error bounds for Riemannian flow matching, with explicit iteration complexities on hyperspheres and Hadamard symmetric spaces. The symmetric-space assumption allows us to present a clean explicit result, and is not essential to the overall approach. While $\nabla R = 0$ simplifies the regularity analysis, a similar approach could extend to more general Hadamard manifolds under suitable bounds on ∇R and $\nabla^2 R$. On the other hand, for hyperspheres, the key regularity question is whether the population vector field remains sufficiently smooth in the presence of a cut locus. Our analysis shows that the Jacobian factor compensates for the nonsmooth cut-locus indicator, yielding a C^2 population vector field and providing the first explicit treatment of this issue for Riemannian flow matching.

Future work includes: (i) extending this regularity analysis to general compact manifolds with more complicated cut loci; and (ii) investigating whether higher-order geometric integrators can improve the $O(h)$ discretization error of the Riemannian Euler scheme.

Acknowledgments and Disclosure of Funding

Research of Krishnakumar Balasubramanian is supported in part by National Science Foundation (NSF) grant DMS-2413426. Research of Shiqian Ma is supported in part by ONR Grant N00014-24-1-2705 and NSF Grants CCF-2626366 and ECCS-2639042.

References

[1] Foivos Alimisis, Antonio Orvieto, Gary Bécigneul, and Aurelien Lucchi. A continuous-time perspective for modeling acceleration in Riemannian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1297–1307. PMLR, 2020.

[2] Vansh Bansal, Saptarshi Roy, Purnamrita Sarkar, and Alessandro Rinaldo. On the Wasserstein Convergence and Straightness of Rectified Flow. *arXiv preprint arXiv:2410.14949*, 2024.

[3] J Benton, G Deligiannidis, and A Doucet. Error bounds for flow matching methods. *Transactions on Machine Learning Research*, 2024.

[4] Jeff Cheeger, David G Ebin, and David Gregory Ebin. *Comparison theorems in Riemannian geometry*, volume 9. North-Holland publishing company Amsterdam, 1975.

[5] Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In *The Twelfth International Conference on Learning Representations*, 2024.

[6] Austin Henry Cheng, Alston Lo, Kin Long Kelvin Lee, Santiago Miret, and Alan Aspuru-Guzik. Stiefel Flow Matching for Moment-Constrained Structure Elucidation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[7] 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.

[8] Antoine Collas, Ce Ju, Nicolas Salvy, and Bertrand Thirion. Riemannian flow matching for brain connectivity matrices via pullback geometry. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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

[10] Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. *Advances in neural information processing systems*, 35:2406–2422, 2022.

[11] Khashayar Gatmiry and Santosh S Vempala. Convergence of the Riemannian Langevin algorithm. *arXiv preprint arXiv:2204.10818*, 2022.

[12] Yunrui Guan, Krishna Balasubramanian, and Shiqian Ma. Riemannian Proximal Sampler for High-accuracy Sampling on Manifolds. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[13] Yunrui Guan, Krishnakumar Balasubramanian, and Shiqian Ma. Mirror flow matching with heavy-tailed priors for generative modeling on convex domains. In *The Fourteenth International Conference on Learning Representations*, 2026.

[14] Hiroshi Hirai, Harold Nieuwboer, and Michael Walter. Interior-point methods on manifolds: theory and applications. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 2021–2030. IEEE, 2023.

[15] 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.

[16] Daniel Zhengyu Huang, Jiaoyang Huang, and Zhengjiang Lin. Convergence analysis of probability flow ODE for score-based generative models. *IEEE Transactions on Information Theory*, 2025.

[17] Shoshichi Kobayashi. *Transformation groups in differential geometry*. Springer, 1972.

[18] Lingkai Kong and Molei Tao. Convergence of kinetic Langevin Monte Carlo on Lie groups. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 3011–3063. PMLR, 2024.

[19] Serge Lang. *Differential and Riemannian manifolds*, volume 160. Springer Science & Business Media, 2012.

[20] John M. Lee. *Introduction to Smooth Manifolds*. Springer, New York, 2012.

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

[22] Mario Lezcano-Casado. Curvature-dependant global convergence rates for optimization on manifolds of bounded geometry. *arXiv preprint arXiv:2008.02517*, 2020.

[23] Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. A sharp convergence theory for the probability flow odes of diffusion models. *arXiv preprint arXiv:2408.02320*, 2024.

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

[25] Runjia Li, Qiwei Di, and Quanquan Gu. Unified convergence analysis for score-based diffusion models with deterministic samplers. In *The Thirteenth International Conference on Learning Representations*, 2025.

[26] Yunchen Li, Zhou Yu, Gaoqi He, Yunhang Shen, Ke Li, Xing Sun, and Shaohui Lin. SPD-DDPM: Denoising diffusion probabilistic models in the symmetric positive definite space. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 13709–13717, 2024.

[27] Yuhao Liu, Rui Hu, Yu Chen, and Longbo Huang. Finite-Time Convergence Analysis of ODE-based Generative Models for Stochastic Interpolants. *arXiv preprint arXiv:2508.07333*, 2025.

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

[29] Xiaoshan Luo, Zhenyu Wang, Qingchang Wang, Xuechen Shao, Jian Lv, Lei Wang, Yanchao Wang, and Yanming Ma. Crystalflow: a flow-based generative model for crystalline materials. *Nature Communications*, 16(1):9267, 2025.

[30] Jerrold E. Marsden, Tudor Ratiu, and Ralph Abraham. *Manifolds, Tensor Analysis, and Applications*. Applied Mathematical Sciences. Springer, 3rd edition, 2002.

[31] Emile Mathieu and Maximilian Nickel. Riemannian continuous normalizing flows. *Advances in neural information processing systems*, 33:2503–2515, 2020.

[32] Benjamin Kurt Miller, Ricky T. Q. Chen, Anuroop Sriram, and Brandon M Wood. FlowMM: Generating materials with Riemannian flow matching. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 35664–35686, 2024.

[33] Saptarshi Roy, Alessandro Rinaldo, and Purnamrita Sarkar. Low-Dimensional Adaptation of Rectified Flow: A New Perspective through the Lens of Diffusion and Stochastic Localization. *arXiv preprint arXiv:2601.15500*, 2026.

[34] 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.

[35] Anuroop Sriram, Benjamin K Miller, Ricky T Chen, and Brandon M Wood. FlowLLM: Flow matching for material generation with large language models as base distributions. *Advances in Neural Information Processing Systems*, 37:46025–46046, 2024.

[36] Maojiang Su, Jerry Yao-Chieh Hu, Sophia Pi, and Han Liu. On flow matching kl divergence. *arXiv preprint arXiv:2511.05480*, 2025.

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

[38] Zhengchao Wan, Qingsong Wang, Gal Mishne, and Yusu Wang. Elucidating Flow Matching ODE Dynamics via Data Geometry and Denoisers. In *Forty-second International Conference on Machine Learning*, 2025.

[39] Jiawen Wu, Bingguang Chen, Yuyi Zhou, Qi Meng, Rongchan Zhu, and Zhi-Ming Ma. Riemannian Neural Geodesic Interpolant. *arXiv preprint arXiv:2504.15736*, 2025.

[40] Xingyu Xu, Ziyi Zhang, Yorie Nakahira, Guannan Qu, and Yuejie Chi. Polynomial Convergence of Riemannian Diffusion Models. *arXiv preprint arXiv:2601.02499*, 2026.

[41] Konstantin Yakovlev, Anna Markovich, and Nikita Puchkin. Implicit score matching meets denoising score matching: improved rates of convergence and log-density hessian estimation. *arXiv preprint arXiv:2512.24378*, 2025.

[42] Angxiao Yue, Zichong Wang, and Hongteng Xu. ReQFlow: Rectified quaternion flow for efficient and high-quality protein backbone generation. In *Forty-second International Conference on Machine Learning*, 2025.

[43] Zixuan Zhang, Kaixuan Huang, Tuo Zhao, Mengdi Wang, and Minshuo Chen. Diffusion model for manifold data: Score decomposition, curvature, and statistical complexity. *arXiv preprint arXiv:2603.20645*, 2026.

[44] Zhengyu Zhou and Weiwei Liu. An Error Analysis of Flow Matching for Deep Generative Modeling. In *Forty-second International Conference on Machine Learning*, 2025.

A Additional Background for Riemannian Geometry

A.1 Distances used

Definition 8 (Total variation distance and W_1 distance)**.** *Let (M, g) be a Riemannian manifold, let d denote the geodesic distance induced by g, and let ρ_1, ρ_2 be probability measures on the Borel σ-algebra $\mathcal{B}(M)$.*

- *The total variation distance between ρ_1 and ρ_2 is*

$$\mathrm{TV}\left(\rho_1, \rho_2\right) \;:=\; \sup_{A \in \mathcal{B}(M)} \left|\rho_1(A) - \rho_2(A)\right|.$$

If ρ_1 and ρ_2 are absolutely continuous with respect to the Riemannian volume measure dV_g, with densities $p_1 = \frac{d\rho_1}{dV_g}$ and $p_2 = \frac{d\rho_2}{dV_g}$, then

$$\mathrm{TV}\left(\rho_1, \rho_2\right) \;=\; \frac{1}{2}\int_M \left|p_1(x) - p_2(x)\right| dV_g(x) \;=\; \int_{\{p_1 > p_2\}} \big(p_1(x) - p_2(x)\big)\, dV_g(x).$$

- *The 1-Wasserstein distance (induced by d) is defined for measures with finite first moment (e.g., $\int_M d(x_0, x)\, d\rho_i(x) < \infty$ for some $x_0 \in M$ and $i \in \{1, 2\}$) by*

$$W_1(\rho_1, \rho_2) \;:=\; \inf_{\gamma \in \Pi(\rho_1, \rho_2)} \int_{M \times M} d(x, y)\, d\gamma(x, y),$$

where $\Pi(\rho_1, \rho_2)$ denotes the set of couplings of (ρ_1, ρ_2), i.e., probability measures γ on $M \times M$ whose first and second marginals are ρ_1 and ρ_2, respectively.

A.2 Comparison Theorem

Comparison theorems [4] will be used frequently throughout the Appendix. For example, we used Rauch theorem in Appendix C.1, which is a key step in proving Lemma 5; we also use comparison theorems frequently in Appendix E to verify Assumption 2.

Here we provide a brief introduction to comparison theory, focusing on intuition. Let $k \in \mathbb{R}$ be some constant, and assume we have functions $f, g \geq 0$ defined on $[0, t]$. If both f, g share the same initial value $f(0) = g(0)$, and we further assume $f'(0) \leq g'(0)$. If we know, for second order ODEs,

$f'' + \kappa f \leq g'' + \kappa g$, then we can conclude, at least locally, $f \leq g$. For more details, see, for example [22, Lemma 4.8]. Such an observation suggests a more general comparison principle, which is the key idea behind comparison theorems in Riemannian geometry. For example, we have the following theorem.

Theorem 9 (Proposition 4.9 in [22])**.** *Let M be a Riemannian manifold with bounded sectional curvature satisfying $K_{\min} \leq \mathrm{Sec} \leq K_{\max}$. Let $\gamma : [0, r] \to M$ be a geodesic, and let X, Y be vector fields along γ with $X, Y \perp \gamma'$ satisfying the following ODE ($X'$is covariant derivative of X along γ):*

$$X'' + R(X, \gamma')\gamma' = Y, \qquad X(0) = 0, X'(0) = 0.$$

Assume that there exists a continuous function η that upper bounds Y: $\|Y\| \leq \eta$ on $[0, r]$. Then ρ defined as the solution of $\rho'' + K_{\min}\rho = \eta, \rho(0) = 0, \rho'(0) = 0$ upper bounds X: $\|X\| \leq \rho$.

To summarize, using the comparison principle for ODE, we can make use of curvature information to establish bounds for certain vector fields defined on a Riemannian manifold, as long as the vector field satisfies some specific ODE. For a comprehensive discussion on comparison theorems, see for example [21] and [4].

A.3 Riemannian Submanifolds

We briefly introduce some results of Riemannian submanifolds following [21], and these results will be used in proving Lemma 49. Let $N \subseteq M$ be a Riemannian submanifold of M, with induced metric. To avoid ambiguity, we denote ∇^M to be the connection on M, and ∇^N to be the connection on N. Let X, Y be vector fields on N. We can extend them to vector fields on M, and the covariant derivative $(\nabla^M)_X Y$ can be decomposed as

$$(\nabla^M)_X Y = ((\nabla^M)_X Y)^\perp + ((\nabla^M)_X Y)^\parallel.$$

The normal component $((\nabla^M)_X Y)^\perp$ defines the second fundamental form, which is a map from $\mathfrak{X}(N) \times \mathfrak{X}(N)$ to a section of the normal bundle of N, formally defined as

$$\mathrm{II}(X, Y) := ((\nabla^M)_X Y)^\perp.$$

Gauss formula [21, Theorem 8.2] states that, for X, Y being vector fields on N and extended arbitrarily to M, then

$$(\nabla^M)_X Y = (\nabla^N)_X Y + \mathrm{II}(X, Y).$$

A closely related terminology is totally geodesic manifold. N is said to be totally geodesic if every geodesic in N is also a geodesic in M. Furthermore, by [21, Proposition 8.12], we know the following are equivalent:

- N is a totally geodesic submanifold of M.
- The second fundamental form of N vanishes identically.
- Every geodesic in N is also a geodesic in M.

B Proof of Main Theorem

This Section is organized as follows. We first prove Lemma 4 in Appendix B.1, which provides a way to control the propagation of TV distance along ODE simulation:

$$\frac{\partial \,\mathrm{TV}\,(X_t, Y_t)}{\partial t} = \int_{\Omega_t} p(t, x)\,(\mathrm{Div}\,(\tilde{v}(x, t) - v(x, t)) + \langle \mathrm{Grad}\,\log p(t, x), \tilde{v}(x, t) - v(x, t)\rangle)\, dV_g(x).$$

Note that the time derivative of TV distance consists of two parts:

- We control the "velocity vector" term $p(t, x)\langle \mathrm{Grad}\,\log p(t, x), \tilde{v}(x, t) - v(x, t)\rangle$ in Appendix B.2.
- We control the "divergence term" $p(t, x)\,\mathrm{Div}\,(\tilde{v}(x, t) - v(x, t))$ in Appendix B.3.

Moreover, recall that the well-definedness of $\tilde{v}$ depends on the invertibility of $F_{t,h}(x) := \text{Exp}_x(h\hat{v}(t,x))$, which we justify in Appendix B.4. The proof of the main result (sampling error bound in Theorem 1) is presented in Appendix B.5. Finally, we present proofs for iteration complexity on the hypersphere and SPD manifold in Appendix B.6 and B.7, respectively.

Appendix C contains auxiliary results needed in the proof of Theorem 1. Finally, Appendix D and E proves the required regularity results needed for proving Propositions 3 and 2.

B.1 Proof for Lemma 4

We start with the following result.

Theorem 10 (Transport Theorem [30], Theorem 8.1.12)**.** *Let (M, μ) be a volume manifold and X be a vector field on M with flow F_t. For smooth function f defined on $\mathcal{F}(M \times \mathbb{R})$, let $f_t(m) = f(m,t)$. We have that for any open set $U \subseteq M$,*

$$\frac{d}{dt}\int_{F_t(U)} f_t \mu = \int_{F_t(U)} \left(\frac{\partial f}{\partial t} + \textit{Div}_{\,\mu}(f_t X)\right)\mu.$$

Proof. [Proof of Lemma 4] We can write

$$\text{TV}\left(X_t, Y_t\right) = \int_{\Omega_t} p(t,x) - q(t,x) dV_g(x).$$

By transport Theorem,

$$\begin{aligned}
\frac{\partial \,\text{TV}\left(X_t, Y_t\right)}{\partial t} &= \frac{d}{dt}\int_{\Omega_t} p(t,x) - q(t,x) dV_g(x) \\
&= \int_{\Omega_t} \left(\frac{\partial (p(t,x) - q(t,x))}{\partial t} + \text{Div}\left((p(t,x) - q(t,x))X\right)\right) dV_g(x) \\
&= \int_{\Omega_t} \frac{\partial (p(t,x) - q(t,x))}{\partial t} dV_g(x) + \int_{\partial\Omega_t} (p(t,x) - q(t,x))\langle X, n\rangle dV_{\hat{g}}(x) \\
&= \int_{\Omega_t} \frac{\partial (p(t,x) - q(t,x))}{\partial t} dV_g(x) + \int_{\partial\Omega_t} 0 \times \langle X, n\rangle dV_{\hat{g}}(x) \\
&= \int_{\Omega_t} \frac{\partial (p(t,x) - q(t,x))}{\partial t} dV_g(x).
\end{aligned}$$

The following continuity equation was proved in [39, Theorem 2]. We provide a proof for completeness. For any test function φ that is smooth (and of bounded support if the manifold is non-compact and does not have boundary), we have

$$\begin{aligned}
\int_M \varphi(x)\frac{\partial}{\partial t} p(t,x) dV_g(x) = \frac{d}{dt}\int_M \varphi(x) p(t,x) dV_g(x) &= \frac{d}{dt}\mathbb{E}[\varphi(x_t)] = \mathbb{E}[\nabla\varphi(x_t) \circ \frac{d}{dt}x_t] \\
&= \int_M \langle \text{Grad}\ \varphi(x), v(x,t)\rangle p(t,x) dV_g(x) \\
&= -\int_M \varphi(x)\,\text{Div}\,(v(x,t)p(t,x)) dV_g(x).
\end{aligned}$$

Hence we conclude that

$$\frac{\partial}{\partial t} p(t,x) = -\,\text{Div}\,(v(x,t)p(t,x)).$$

Therefore

$$\begin{aligned}
\frac{\partial \,\text{TV}\left(X_t, Y_t\right)}{\partial t} &= \int_{\Omega_t} \frac{\partial (p(t,x) - q(t,x))}{\partial t} dV_g(x) \\
&= \int_{\Omega_t} -\,\text{Div}\,(v(x,t)p(t,x)) + \text{Div}\,(\tilde{v}(x,t)q(t,x)) dV_g(x)
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\partial\Omega_t} -p(t,x)\langle v(x,t), n\rangle + q(t,x)\langle \tilde{v}(x,t), n\rangle dV_{\hat{g}}(x) \\
&= \int_{\partial\Omega_t} p(t,x)\langle \tilde{v}(x,t) - v(x,t), n\rangle dV_{\hat{g}}(x) \\
&= \int_{\Omega_t} \text{Div}\left((\tilde{v}(x,t) - v(x,t))p(t,x)\right) dV_g(x) \\
&= \int_{\Omega_t} p(t,x)\,\text{Div}\left(\tilde{v}(x,t) - v(x,t)\right) + \langle \text{Grad}\ p(t,x), \tilde{v}(x,t) - v(x,t)\rangle dV_g(x) \\
&= \int_{\Omega_t} p(t,x)\left(\text{Div}\left(\tilde{v}(x,t) - v(x,t)\right) + \langle \text{Grad}\ \log p(t,x), \tilde{v}(x,t) - v(x,t)\rangle\right) dV_g(x),
\end{aligned}
$$

where observe that $p = q$ on $\partial\Omega_t$, and the last equality is due to the product rule of divergence. Note that Grad $p(t,x) = p(t,x)$ Grad $\log p(t,x)$, thereby concluding the proof. □

B.2 Velocity Vector Term

In this section, we bound the velocity vector term.

Lemma 11. *Under Assumption 2, 3 and 4, we can bound*

$$
\begin{aligned}
&\mathbb{E}[\|\textit{Grad}\,\log p(t,x)\| \cdot \|\tilde{v}(x,t) - v(x,t)\|] \\
\leq &\varepsilon\sqrt{3L_t^{\textit{score}}} + \left(L_t^{\textit{score}}(6(t-t_k)^2\left((L_t^{\hat{v},x})^2(L_t^{\hat{v}})^2 + (L_t^{\hat{v},x})^2(L_t^{v})^2 + (L_t^{v,t})^2\right)\right)^{\frac{1}{2}}.
\end{aligned}
$$

Proof. Throughout this proof, $P_{X_s}^{X_{t_k}}$ and its inverse denote parallel transport along the population ODE trajectory, and $P_{X_{t\to t_k}}^{X_t}$ denote parallel transport along the geodesic. Using triangle inequality, we can write

$$
\begin{aligned}
&\|\tilde{v}(t,X_t) - v(t,X_t)\| \\
\leq &\|P_{X_t}^{X_{t_k}}\tilde{v}(t,X_t) - \hat{v}(t_k,X_{t_k})\| + \|\hat{v}(t_k,X_{t_k}) - v(t_k,X_{t_k})\| + \|v(t_k,X_{t_k}) - P_{X_t}^{X_{t_k}}v(t,X_t)\|.
\end{aligned}
$$

Hence using Cauchy-Schwarz,

$$
\begin{aligned}
&\mathbb{E}[\|\tilde{v}(t,X_t) - v(t,X_t)\|^2] \\
\leq &\mathbb{E}[3\|P_{X_t}^{X_{t_k}}\tilde{v}(t,X_t) - \hat{v}(t_k,X_{t_k})\|^2 + 3\|\hat{v}(t_k,X_{t_k}) - v(t_k,X_{t_k})\|^2 + 3\|v(t_k,X_{t_k}) - P_{X_t}^{X_{t_k}}v(t,X_t)\|^2].
\end{aligned}
$$

Denote $X_{t\to t_k} = F_{t_k,t-t_k}^{-1}(X_t)$.

For the first term, we have

$$
\begin{aligned}
&\mathbb{E}\|P_{X_t}^{X_{t_k}}\tilde{v}(t,X_t) - \hat{v}(t_k,X_{t_k})\|^2 \\
= &\mathbb{E}\|P_{X_t}^{X_{t_k}}P_{X_{t\to t_k}}^{X_t}\hat{v}(t_k,X_{t\to t_k}) - \hat{v}(t_k,X_{t_k})\|^2 = \mathbb{E}\|P_{X_{t\to t_k}}^{X_t}\hat{v}(t_k,X_{t\to t_k}) - P_{X_{t_k}}^{X_t}\hat{v}(t_k,X_{t_k})\|^2 \\
\leq &2\mathbb{E}[\|P_{X_{t\to t_k}}^{X_t}\hat{v}(t_k,X_{t\to t_k}) - \hat{v}(t_k,X_t)\|^2 + \|\hat{v}(t_k,X_t) - P_{X_{t_k}}^{X_t}\hat{v}(t_k,X_{t_k})\|^2] \\
\leq &2(L_t^{\hat{v},x})^2\mathbb{E}[d(X_{t\to t_k},X_t)^2] + 2(L_t^{\hat{v},x})^2\mathbb{E}[(\int_{t_k}^t \|v(s,X_s)\|ds)^2] \\
= &2(L_t^{\hat{v},x})^2\mathbb{E}\left[d(F_{t_k,t-t_k}^{-1}(X_t),X_t)^2 + (\int_{t_k}^t \|v(s,X_s)\|ds)^2\right],
\end{aligned}
$$

where we used the fact that parallel transport preserve norm.

Note that X_t is the trajectory of X at time t. Starting from $X_{t\to t_k}$, we have

$$
\text{Exp}_{X_{t\to t_k}}((t-t_k)\hat{v}(t_k,X_{t\to t_k})) = X_t.
$$

Hence

$$
\mathbb{E}[d(F_{t_k,t-t_k}^{-1}(X_t),X_t)^2] = \mathbb{E}[(t-t_k)^2\|\hat{v}(t_k,X_{t\to t_k})\|^2] \leq (t-t_k)^2(L_t^{\hat{v}})^2.
$$

Also, the ODE trajectory has length $\int_{t_k}^{t} \|v(s, X_s)\| ds$, hence

$$\begin{aligned}\mathbb{E}[d(X_t, X_{t_k})^2] &\le \mathbb{E}[(\int_{t_k}^{t} \|v(s, X_s)\| ds)^2] \le (t - t_k)\mathbb{E}[\int_{t_k}^{t} \|v(s, X_s)\|^2 ds] \\ &= (t - t_k) \int_{t_k}^{t} \mathbb{E}[\|v(s, X_s)\|^2] ds \le (t - t_k)^2 (L_t^v)^2.\end{aligned}$$

Now, for the second term, we have

$$\mathbb{E}\|\hat{v}(t_k, X_{t_k}) - v(t_k, X_{t_k})\|^2 \le \varepsilon^2.$$

Finally, for the third term, by swapping the order of the derivative and parallel transport, we get

$$\|v(t_k, X_{t_k}) - P_{X_t}^{X_{t_k}} v(t, X_t)\| \le \int_{t_k}^{t} \left\| \frac{d}{ds} P_{X_s}^{X_{t_k}} v(s, X_s) \right\| ds = \int_{t_k}^{t} \left\| \partial_s v(s, X_s) + \nabla_{\dot{X}_s} v(s, X_s) \right\| ds.$$

Therefore,

$$\mathbb{E}\|v(t_k, X_{t_k}) - P_{X_t}^{X_{t_k}} v(t, X_t)\|^2 \le (t - t_k) \int_{t_k}^{t} \mathbb{E}\left\| \partial_s v(s, X_s) + \nabla_{\dot{X}_s} v(s, X_s) \right\|^2 ds.$$

Notice that

$$\mathbb{E}\left\| \partial_s v(s, X_s) + \nabla_{\dot{X}_s} v(s, X_s) \right\|^2 = \mathbb{E}\left\| \frac{D}{ds} v(s, X_s) \right\|^2 \le (L_s^{v,t})^2,$$

so we have

$$\begin{aligned}\mathbb{E}\|v(t_k, X_{t_k}) - P_{X_t}^{X_{t_k}} v(t, X_t)\|^2 \le& (t - t_k) \int_{t_k}^{t} \mathbb{E}\left\| \partial_s v(s, X_s) + \nabla_{\dot{X}_s} v(s, X_s) \right\|^2 ds \\ \le& (t - t_k) \int_{t_k}^{t} (L_s^{v,t})^2 ds \le (t - t_k)^2 ((L_t^{v,t})^2).\end{aligned}$$

Putting the above estimates together, we obtain

$$\begin{aligned}&\mathbb{E}[\|\tilde{v}(t, X_t) - v(t, X_t)\|^2] \\ \le& \mathbb{E}[3\|P_{X_t}^{X_{t_k}} \tilde{v}(t, X_t) - \hat{v}(t_k, X_{t_k})\|^2 + 3\|\hat{v}(t_k, X_{t_k}) - v(t_k, X_{t_k})\|^2 + 3\|v(t_k, X_{t_k}) - P_{X_t}^{X_{t_k}} v(t, X_t)\|^2] \\ \le& 6(L_t^{\hat{v},x})^2 \left((t - t_k)^2 (L_t^{\hat{v}})^2 + (t - t_k)^2 (L_t^v)^2 \right) + 3\varepsilon^2 + 6(t - t_k)^2 (L_t^{v,t})^2 \\ =& 6(t - t_k)^2 \left((L_t^{\hat{v},x})^2 (L_t^{\hat{v}})^2 + (L_t^{\hat{v},x})^2 (L_t^v)^2 + (L_t^{v,t})^2 \right) + 3\varepsilon^2.\end{aligned}$$

Hence

$$\begin{aligned}&\mathbb{E}[\|\text{ Grad } \log p(t, x)\| \cdot \|\tilde{v}(x, t) - v(x, t)\|] \\ \le& \mathbb{E}[\|\text{ Grad } \log p(t, x)\|^2]^{\frac{1}{2}} \mathbb{E}[\|\tilde{v}(x, t) - v(x, t)\|^2]^{\frac{1}{2}} \\ \le& \left(L_t^{\text{score}} (6(t - t_k)^2 \left((L_t^{\hat{v},x})^2 (L_t^{\hat{v}})^2 + (L_t^{\hat{v},x})^2 (L_t^v)^2 + (L_t^{v,t})^2 \right) + 3\varepsilon^2) \right)^{\frac{1}{2}} \\ \le& \varepsilon \sqrt{3 L_t^{\text{score}}} + \left(L_t^{\text{score}} (6(t - t_k)^2 \left((L_t^{\hat{v},x})^2 (L_t^{\hat{v}})^2 + (L_t^{\hat{v},x})^2 (L_t^v)^2 + (L_t^{v,t})^2 \right)) \right)^{\frac{1}{2}}.\end{aligned}$$

□

B.3 Divergence Term

In this section, we bound the divergence term.

Denote $z = F_{t_k, t - t_k}^{-1}(x)$. We know $\text{Exp}_z((t - t_k)\hat{v}(t_k, z)) = x$. Recall we denote

$$\tilde{v}(t, x) = P_{F_{t_k, t - t_k}^{-1}(x)}^{x} \hat{v}(t_k, F_{t_k, t - t_k}^{-1}(x)),$$

by simply,

$$\tilde{v}(x) = P_z^{\text{Exp}_z(h\hat{v}(z))} \hat{v}(z)$$

for notational simplicity.

Lemma 12. *Under Assumption 2, 3 and 4, and the stepsize conditions of Lemma 18, we have*

$$\begin{aligned}&\mathbb{E}_{p(t,x)}[|\,\mathit{Div}\,(\tilde{v}(t,X_t)-v(t,X_t))|]\\ \le&3\varepsilon+(t-t_k)\Big(L_t^{\mathit{Div}\ \hat{v},x}(L_t^{\hat{v}}+L_t^{v})+L_t^{\mathit{Div},t}+5L_R(L_t^{\hat{v}})^2d+2(L_t^{\hat{v},x})^2\Big).\end{aligned}$$

Proof.

By Lemma 18, we have

$$\begin{aligned}&\mathbb{E}_{p(t,x)}[|\,\mathrm{Div}\,(\tilde{v}(t,X_t)-v(t,X_t))|]\\ \le&\mathbb{E}_{p(t,x)}[|\mathrm{Div}\ \hat{v}(t_k,z)-\mathrm{Div}\ v(t,X_t)|]+(t-t_k)(2(L_t^{\hat{v},x})^2+5dL_R\mathbb{E}_{p(t,x)}[\|\hat{v}(t_k,z)\|^2])\\ \le&\mathbb{E}_{p(t,x)}[|\mathrm{Div}\ \hat{v}(t_k,X_{t_k})-\mathrm{Div}\ v(t_k,X_{t_k})|]+\mathbb{E}_{p(t,x)}[|\mathrm{Div}\ \hat{v}(t_k,z)-\mathrm{Div}\ \hat{v}(t_k,X_{t_k})|]\\ &+\mathbb{E}_{p(t,x)}[|\mathrm{Div}\ v(t_k,X_{t_k})-\mathrm{Div}\ v(t,X_t)|]+(t-t_k)(2(L_t^{\hat{v},x})^2+5dL_R(L_t^{\hat{v}})^2).\end{aligned}$$

Recall

$$\begin{aligned}\mathbb{E}[d(F^{-1}_{t_k,t-t_k}(X_t),X_t)^2]\le&(t-t_k)^2(L_t^{\hat{v}})^2,\\ \mathbb{E}[d(X_t,X_{t_k})^2]\le&(t-t_k)^2(L_t^{v})^2.\end{aligned}$$

For the first term, since $X_{t_k}\sim p(t_k,\cdot)$, Assumption 4 gives

$$\mathbb{E}_{p(t,x)}[|\mathrm{Div}\ \hat{v}(t_k,X_{t_k})-\mathrm{Div}\ v(t_k,X_{t_k})|]\le\varepsilon.$$

For the second term, by the pointwise bound on $\|\,\mathrm{Grad}_x\,\mathrm{Div}\ \hat{v}(t_k,\cdot)\|$, the function $\mathrm{Div}\ \hat{v}(t_k,\cdot)$ is Lipschitz:

$$\big|\,\mathrm{Div}\ \hat{v}(t_k,z)-\mathrm{Div}\ \hat{v}(t_k,X_{t_k})\big|\le L_t^{\mathrm{Div}\ \hat{v},x}\,d(z,X_{t_k}).$$

Finally, using $d(z,X_{t_k})\le d(z,X_t)+d(X_t,X_{t_k})$ and the bounds

$$\mathbb{E}[d(z,X_t)]\le(t-t_k)L_t^{\hat{v}},\qquad\mathbb{E}[d(X_t,X_{t_k})]\le(t-t_k)L_t^{v},$$

we obtain

$$\mathbb{E}_{p(t,x)}\big[\big|\,\mathrm{Div}\ \hat{v}(t_k,z)-\mathrm{Div}\ \hat{v}(t_k,X_{t_k})\big|\big]\le(t-t_k)L_t^{\mathrm{Div}\ \hat{v},x}(L_t^{\hat{v}}+L_t^{v}).$$

For the third term, by the fundamental theorem of calculus along the curve, we have

$$\mathrm{Div}\ v(t,X_t)-\mathrm{Div}\ v(t_k,X_{t_k})=\int_{t_k}^{t}(\partial_s\,\mathrm{Div}\ v(s,X_s)+\langle\mathrm{Grad}_x\,\mathrm{Div}\ v(s,X_s),\partial_sX_s\rangle)\,ds.$$

Since $\partial_sX_s=v(s,X_s)$,

$$\big|\,\mathrm{Div}\ v(t,X_t)-\mathrm{Div}\ v(t_k,X_{t_k})\big|\le\int_{t_k}^{t}|\partial_s(\mathrm{Div}\ v)(s,X_s)+\langle\mathrm{Grad}_x\,\mathrm{Div}\ v(s,X_s),v(s,X_s)\rangle|\,ds.$$

Hence, using $X_s\sim p(s,\cdot)$, we obtain

$$\begin{aligned}&\mathbb{E}_{p(t,x)}\big[\big|\,\mathrm{Div}\ v(t,X_t)-\mathrm{Div}\ v(t_k,X_{t_k})\big|\big]\\ \le&\mathbb{E}_{p(t,x)}\Big[\int_{t_k}^{t}|\partial_s(\mathrm{Div}\ v)(s,X_s)+\langle\mathrm{Grad}_x\,\mathrm{Div}\ v(s,X_s),v(s,X_s)\rangle|\,ds\Big]\\ \le&\int_{t_k}^{t}\mathbb{E}_{p(s,x)}\Big[\Big|\frac{d}{ds}\,\mathrm{Div}\ v(s,X_s)\Big|\Big]ds\le(t-t_k)(L_t^{\mathrm{Div},t}).\end{aligned}$$

Putting together, the above estimates, we obtain

$$\begin{aligned}&\mathbb{E}_{p(t,x)}[|\,\mathrm{Div}\,(\tilde{v}(t,X_t)-v(t,X_t))|]\\ \le&\mathbb{E}_{p(t,x)}[|\mathrm{Div}\ \hat{v}(t_k,X_{t_k})-\mathrm{Div}\ v(t_k,X_{t_k})|]+\mathbb{E}_{p(t,x)}[|\mathrm{Div}\ \hat{v}(t_k,z)-\mathrm{Div}\ \hat{v}(t_k,X_{t_k})|]\\ &+\mathbb{E}_{p(t,x)}[|\mathrm{Div}\ v(t_k,X_{t_k})-\mathrm{Div}\ v(t,X_t)|]+5L_R(t-t_k)(L_t^{\hat{v}})^2d+2(t-t_k)(L_t^{\hat{v},x})^2\\ \le&3\varepsilon+L_t^{\mathrm{Div}\ \hat{v},x}(t-t_k)(L_t^{\hat{v}}+L_t^{v})+(t-t_k)(L_t^{\mathrm{Div},t})+5L_R(t-t_k)(L_t^{\hat{v}})^2d+2(t-t_k)(L_t^{\hat{v},x})^2\\ \le&3\varepsilon+(t-t_k)\Big(L_t^{\mathrm{Div}\ \hat{v},x}(L_t^{\hat{v}}+L_t^{v})+L_t^{\mathrm{Div},t}+5L_R(L_t^{\hat{v}})^2d+2(L_t^{\hat{v},x})^2\Big).\end{aligned}$$

□

B.4 Results on Riemannian manifolds

Note that in Lemma 4, we do not require vector fields $v, \tilde{v}$ to be the flow matching vector field. To apply the Lemma for analyzing the discretization scheme, we will set v to be the true vector field for flow matching, and $\tilde{v}$ to be corresponds to the learned vector field.

But however, there is a discrepency between continuous time ODE $dY_t = \tilde{v}(t, Y_t)dt$ and the Euler discretization scheme. Hence, to apply the Lemma to discretization (the actual method in Algorithm 1), we need to define a continuous time interpolation.

Define F as $F_{t,h}(x) := \text{Exp}_x(h\hat{v}(t, x))$. Note that

$$F_{t_k, t-t_k}(x_k) = \text{Exp}_{x_k}((t - t_k)\hat{v}(t_k, x_k))$$

corresponds to the continuous time interpolation of Euler discretization. Then we are able to define a interpolation vector field: we want to define

$$dY_t = P^{Y_t}_{Y_{t_k}} \hat{v}(t_k, Y_{t_k})dt \stackrel{?}{:=} \tilde{v}(t, Y_t)dt.$$

Here we use the question mark to emphasize that $\tilde{v}$ has not yet been proved to be well defined. We want to write the right hand side as a function of $(t, Y_t) = (t, F_{t_k,t-t_k}(Y_{t_k}))$.

Lemma (Restated Lemma 5)**.** *Let M be simply connected Riemannian manifold that satisfies Assumption 1. Let b be any vector field on M, satisfying $\|b(x)\| \le B, \forall x \in M$. Assume $\|\nabla_v b(x)\| \le L_\nabla \|v\|$. Let $R = Inj\,(M)$. To guarantee $F_{t_k,t-t_k}$ being invertible:*

1. *If $K_{\min} > 0$, we require*

$$h < \min\{\frac{R}{B}, \frac{1}{4L_\nabla}, \sqrt{\frac{3}{4\|b(x)\|^2 L_R(2 + 2L_\nabla \max\{\frac{1}{\sqrt{K_{\min}}}, 1\})}}\}.$$

2. *If $K_{\min} < 0$, we require*

$$h < \min\{\frac{R}{B}, \frac{1}{4L_\nabla}, \sqrt{\frac{3}{4\|b(x)\|^2 L_R(2\frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}} + 4\frac{\cosh(\sqrt{-K_{\min}})-1}{-K_{\min}} L_\nabla)}}\}.$$

3. *If $K_{\min} = 0$, we require*

$$h < \min\{\frac{R}{B}, \frac{1}{4L_\nabla}, \sqrt{\frac{3}{4\|b(x)\|^2 L_R(2 + hL_\nabla)}}\}.$$

Proof. [Proof of Lemma 5] For ease of notation, we use h to denote the time step and b to denote the learned vector field, for the map F. That is, for $h \in \mathbb{R}$ and $b \in \mathfrak{X}(M)$, we write

$$F_h : M \to M, \qquad F_h(x) := \text{Exp}_x\big(hb(x)\big).$$

We first compute the derivative of F. Let $x \in M$ and $v \in T_xM$. Note that $dF_h(x)$ can be viewed as an operator that maps $v \in T_xM$ to $dF_h(x)v \in T_{F_h(x)}M$. We compute $dF_h(x)v$. Let $c(s)$ be a smooth curve s.t. $c(0) = x$ and $c'(0) = v$.

Define a variation through geodesics as

$$\Lambda : (-\epsilon, \epsilon) \times [0, 1] \to M, \qquad \Lambda(s, t) := \text{Exp}_{c(s)}\big(thb(c(s))\big).$$

For each fixed s, the t direction is the geodesic starting at $c(s)$ with initial velocity $hb(c(s))$. In particular, for every s we have

$$F_h(b(c(s))) = \text{Exp}_{c(s)}(hb(c(s))) = \Lambda(s, 1).$$

Now define a vector field J_v along the central geodesic $\gamma(t) := \Lambda(0, t) = \text{Exp}_x\big(thb(x)\big)$ by

$$J_v(t) := \partial_s \Lambda(0, t).$$

Then, by construction (view the left hand side as directional derivative along the curve induced by v),

$$dF_h(x)(v) = \frac{d}{ds} F_h(b(c(s)))\Big|_{s=0} = \partial_s \Lambda(0,1) = J_v(1).$$

This expresses the differential of F_h at x applied to v as the value at $t=1$ of the variation field J_v along the geodesic γ. Note that J_v is a Jacobi field, hence satisfies the Jacobi equation, see [21, Theorem 10.1]. We have initial conditions $J_v(0) = v$, and $D_t J_v(0) = D_t \partial_s \Lambda(0,0) = D_s \partial_t \Lambda(0,0) = h \nabla_v b(x) =: \omega$.

Now we analyze conditions that guarantee the invertibility of F. Define

$$Y(t) = P_{\gamma(t)}^{\gamma(0)} J_v(t).$$

To show that $dF_h(x)$ is invertible, it suffices to show $\inf_v \frac{\|dF_h(x)v\|}{\|v\|} > 0$, equivalently,

$$\|dF_h(x)v\| = \|P_{\gamma(1)}^{\gamma(0)} J_v(1)\| = \|Y(1)\| \geq C\|v\| > 0, \forall v \neq 0$$

for some constant $C > 0$.

Applying Lemma 19 with $c = \gamma$ and $Y = J_v$, we obtain

$$Y'(t) = \frac{d}{dt}\Big(P_{\gamma(t)}^{\gamma(0)} J_v(t)\Big) = P_{\gamma(t)}^{\gamma(0)} D_t J_v(t).$$

Differentiating once more and using the same lemma with $Y = D_t J_v$, we get

$$Y''(t) = \frac{d}{dt}\Big(P_{\gamma(t)}^{\gamma(0)} D_t J_v(t)\Big) = P_{\gamma(t)}^{\gamma(0)} D_t^2 J_v(t).$$

Apply the Jacobi equation $D_t^2 J_v(t) + R(J_v(t), \gamma'(t))\gamma'(t) = 0$ we obtain

$$Y''(t) = P_{\gamma(t)}^{\gamma(0)} D_t^2 J_v(t) = -P_{\gamma(t)}^{\gamma(0)} \left(R(J_v(t), \gamma'(t))\gamma'(t)\right).$$

Using $J_v(t) = P_{\gamma(0)}^{\gamma(t)} Y(t)$, we rewrite the curvature term and obtain

$$Y''(t) + P_{\gamma(t)}^{\gamma(0)} \Big(R\big(P_{\gamma(0)}^{\gamma(t)} Y(t), \gamma'(t)\big)\gamma'(t)\Big) = 0.$$

Now we apply Taylor's theorem with integral remainder entry-wisely to $Y(t)$.

$$\begin{aligned} Y(t) &= Y(0) + tY'(0) + \int_0^t (t-s) Y''(s) ds \\ &= v + th\,\nabla_v b(x) - \int_0^t (t-s) P_{\gamma(s)}^{\gamma(0)} \Big(R\big(P_{\gamma(0)}^{\gamma(s)} Y(s), \gamma'(s)\big)\gamma'(s)\Big) ds. \end{aligned}$$

In particular, for $t = 1$,

$$Y(1) - v - h\nabla_v b(x) = -\int_0^1 (1-s) P_{\gamma(s)}^{\gamma(0)} \Big(R\big(P_{\gamma(0)}^{\gamma(s)} Y(s), \gamma'(s)\big)\gamma'(s)\Big) ds.$$

Since we assumed $\|R(u,v)w\| \leq L_R \|u\|\|v\|\|w\|$,

$$\|R\big(P_{\gamma(0)}^{\gamma(t)} Y(t), \gamma'(t)\big)\gamma'(t)\| \leq L_R \|\gamma'(t)\|^2 \|Y(t)\|.$$

Taking norms, we get

$$\begin{aligned} \|Y(1) - v - h\nabla_v b(x)\| &\leq \int_0^1 (1-s) L_R \|\gamma'(s)\|^2 \|Y(s)\| ds \\ &\leq h^2 \|b(x)\|^2 L_R \int_0^1 \|Y(s)\| ds. \end{aligned}$$

where note $\|\gamma'(s)\| = h\|b(x)\|$. We remark that it suffices to upper bound $\int_0^1 \|Y(s)\| ds$. Then we can simply apply triangle inequality as $\|Y(1)\| \geq \|v + h\nabla_v b(x)\| - \|Y(1) - v - h\nabla_v b(x)\|$.

Now we bound $\int_0^1 \|Y(s)\|ds$ through comparison theory. Since the computation that involves comparison theory is complicated, we summarize them into a separate Lemma, see Lemma 16.

By Lemma 16, when $K_{\min} > 0$, we have

$$\begin{aligned}\int_0^1 \|Y(t)\|dt \leq & \|v^{\|}\| + \|\omega^{\|}\| + \frac{\|\omega^{\perp}\|}{\sqrt{K_{\min}}} + \|v^{\perp}\| \leq 2\|v\| + 2h\|\nabla_v b(x)\| \max\{\frac{1}{\sqrt{K_{\min}}}, 1\} \\ \leq & \|v\|(2 + 2hL_{\nabla} \max\{\frac{1}{\sqrt{K_{\min}}}, 1\}) =: C_+,\end{aligned}$$

where note that we did orthogonal decomposition on v, ω.

$$\|Y(1) - v - h\,\nabla_v b(x)\| \leq h^2 \|b(x)\|^2 L_R C_+.$$

By triangle inequality,

$$\begin{aligned}\|Y(1)\| \geq & \|v + h\nabla_v b(x)\| - \|Y(1) - v - h\nabla_v b(x)\| \\ \geq & \|v\| - h\|\nabla_v b(x)\| - h^2\|b(x)\|^2 L_R C_+ \\ \geq & \|v\| \left(1 - hL_{\nabla} - h^2\|b(x)\|^2 L_R(2 + 2hL_{\nabla} \max\{\frac{1}{\sqrt{K_{\min}}}, 1\})\right).\end{aligned}$$

Clearly $h < 1$ is required, so we only need to find h s.t.

$$\begin{aligned}1 - hL_{\nabla} > & h^2\|b(x)\|^2 L_R(2 + 2L_{\nabla} \max\{\frac{1}{\sqrt{K_{\min}}}, 1\}) \\ > & h^2\|b(x)\|^2 L_R(2 + 2hL_{\nabla} \max\{\frac{1}{\sqrt{K_{\min}}}, 1\}).\end{aligned}$$

If $h \leq \frac{1}{4L_{\nabla}}$, then $1 - hL_{\nabla} \geq \frac{3}{4}$. The value $h \leq \sqrt{\frac{3}{4\|b(x)\|^2 L_R(2+2L_{\nabla} \max\{\frac{1}{\sqrt{K_{\min}}}, 1\})}}$ suffices.

When $K_{\min} < 0$, we have

$$\begin{aligned}\int_0^1 \|Y(t)\|dt \leq & \|v^{\|}\| + \|\omega^{\|}\| + \frac{\cosh(\sqrt{-K_{\min}}) - 1}{-K_{\min}}\|\omega^{\perp}\| + \frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}}\|v^{\perp}\| \\ \leq & \|v\| \left(2\frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}} + 4\frac{\cosh(\sqrt{-K_{\min}}) - 1}{-K_{\min}} hL_{\nabla}\right) =: C_-.\end{aligned}$$

Then we have

$$\|Y(1) - v - h\,\nabla_v b(x)\| \leq h^2 \|b(x)\|^2 L_R C_-.$$

By triangle inequality,

$$\begin{aligned}\|Y(1)\| \geq & \|v + h\nabla_v b(x)\| - \|Y(1) - v - h\nabla_v b(x)\| \\ \geq & \|v\| - h\|\nabla_v b(x)\| - h^2\|b(x)\|^2 L_R C_- \\ \geq & \|v\| \left(1 - hL_{\nabla} - h^2\|b(x)\|^2 L_R(2\frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}} + 4\frac{\cosh(\sqrt{-K_{\min}}) - 1}{-K_{\min}} hL_{\nabla})\right).\end{aligned}$$

Cleraly $h < 1$ is required, so we only need to find h s.t.

$$\begin{aligned}1 - hL_{\nabla} > & h^2\|b(x)\|^2 L_R(2\frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}} + 4\frac{\cosh(\sqrt{-K_{\min}}) - 1}{-K_{\min}} L_{\nabla}) \\ > & h^2\|b(x)\|^2 L_R(2\frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}} + 4\frac{\cosh(\sqrt{-K_{\min}}) - 1}{-K_{\min}} hL_{\nabla}).\end{aligned}$$

If $h \leq \frac{1}{4L_{\nabla}}$, then $1 - hL_{\nabla} \geq \frac{3}{4}$. The value $h \leq \sqrt{\frac{3}{4\|b(x)\|^2 L_R(2\frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}} + 4\frac{\cosh(\sqrt{-K_{\min}})-1}{-K_{\min}} L_{\nabla})}}$ suffices.

When $K_{\min} = 0$, we have

$$\begin{aligned}\int_0^1 \|Y(t)\|dt \leq& \|v^{\|}\| + \frac{1}{2}\|\omega^{\|}\| + \frac{1}{2}\|\omega^{\perp}\| + \|v^{\perp}\| \\ \leq& \|v\|\,(2 + hL_{\nabla}) =: C_{=}.\end{aligned}$$

Then we have

$$\|Y(1) - v - h\,\nabla_v b(x)\| \leq h^2\|b(x)\|^2 L_R C_{=}.$$

By triangle inequality,

$$\begin{aligned}\|Y(1)\| &\geq \|v + h\nabla_v b(x)\| - \|Y(1) - v - h\nabla_v b(x)\| \\ &\geq \|v\| - h\|\nabla_v b(x)\| - h^2\|b(x)\|^2 L_R C_{=} \\ &\geq \|v\|\left(1 - hL_{\nabla} - h^2\|b(x)\|^2 L_R(2 + hL_{\nabla})\right).\end{aligned}$$

Cleraly $h < 1$ is required, so we only need to find h s.t.

$$1 - hL_{\nabla} > h^2\|b(x)\|^2 L_R(2 + L_{\nabla}) > h^2\|b(x)\|^2 L_R(2 + hL_{\nabla}).$$

If $h \leq \frac{1}{4L_{\nabla}}$, then $1 - hL_{\nabla} \geq \frac{3}{4}$. The value $h \leq \sqrt{\frac{3}{4\|b(x)\|^2 L_R(2+hL_{\nabla})}}$ suffices. □

B.5 Main Theorem

Now we prove our main theorem.

Proof. [Proof of Theorem 1] Under the stepsize conditions of Lemma 18, we have that

$$\begin{aligned}&\mathbb{E}[\|\text{ Grad }\log p(t,x)\| \cdot \|\tilde{v}(x,t) - v(x,t)\|] \\ \leq& \varepsilon\sqrt{3L_t^{\text{score}}} + (t - t_k)\Big(6L_t^{\text{score}}\left((L_t^{\hat{v},x})^2(L_t^{\hat{v}})^2 + (L_t^{\hat{v},x})^2(L_t^{v})^2 + (L_t^{v,t})^2 + (L_t^{v})^2(L_t^{v,x})^2\right)\Big)^{\frac{1}{2}},\end{aligned}$$

and

$$\mathbb{E}_{p(t,x)}[|\text{ Div }(\tilde{v}(x,t) - v(x,t))|] \leq 3\varepsilon + (t - t_k)\Big(L_t^{\text{Div }\hat{v},x}(L_t^{\hat{v}} + L_t^{v}) + L_t^{\text{Div },x}L_t^{v} + L_t^{\text{Div },t} + 5L_R(L_t^{\hat{v}})^2 d + 2(L_t^{\hat{v},x})^2\Big).$$

Also recall

$$\begin{aligned}&\text{TV}\,(\pi_T, \hat{\pi}_T) = \text{TV}\,(\pi_{t_N}, \hat{\pi}_{t_N}) \\ \leq& \text{TV}\,(\pi_{t_0}, \hat{\pi}_{t_0}) + \int_{t_0}^{t_N}\mathbb{E}[|\langle\text{Grad }\log p(t,x), \tilde{v}(x,t) - v(x,t)\rangle|]dt + \int_{t_0}^{t_N}\mathbb{E}[|\text{ Div }(\tilde{v}(x,t) - v(x,t))|]dt \\ \leq& \text{TV}\,(\pi_{t_0}, \hat{\pi}_{t_0}) + \int_{t_0}^{t_N}\mathbb{E}[\|\text{ Grad }\log p(t,x)\| \cdot \|\tilde{v}(x,t) - v(x,t)\|]dt + \int_{t_0}^{t_N}\mathbb{E}[|\text{ Div }(\tilde{v}(x,t) - v(x,t))|]dt.\end{aligned}$$

Notice that

$$\int_{t_k}^{t_{k+1}}(t - t_k)dt = (\frac{1}{2}t^2 - tt_k)|_{t_k}^{t_{k+1}} = \frac{1}{2}t_{k+1}^2 - \frac{1}{2}t_k^2 - t_k(t_{k+1} - t_k) = \frac{1}{2}(t_{k+1} - t_k)^2. \quad (9)$$

Hence using constant step size for discretization, we obtain

$$\begin{aligned}&\text{TV}\,(\pi_T, \hat{\pi}_T) \\ \leq& \text{TV}\,(X_{t_0}, Y_{t_0}) + \int_{t_0}^{t_N}\mathbb{E}[\|\text{ Grad }\log p(t,x)\| \cdot \|\tilde{v}(x,t) - v(x,t)\|]dt + \int_{t_0}^{t_N}\mathbb{E}[|\text{ Div }(\tilde{v}(x,t) - v(x,t))|]dt \\ \leq& \text{TV}\,(X_{t_0}, Y_{t_0}) + \frac{1}{2}Nh^2\Big(\Big(6L_t^{\text{score}}\left((L_t^{\hat{v},x})^2(L_t^{\hat{v}})^2 + (L_t^{\hat{v},x})^2(L_t^{v})^2 + (L_t^{v,t})^2 + (L_t^{v})^2(L_t^{v,x})^2\right)\Big)^{\frac{1}{2}} \\ &+ \Big(L_t^{\text{Div }\hat{v},x}(L_t^{\hat{v}} + L_t^{v}) + L_t^{\text{Div },x}L_t^{v} + L_t^{\text{Div },t} + 5L_R(L_t^{\hat{v}})^2 d + 2(L_t^{\hat{v},x})^2\Big)\Big) + Nh(\varepsilon\sqrt{3L_t^{\text{score}}} + 3\varepsilon) \\ &\quad (10) \\ \leq& \text{TV}\,(X_{t_0}, Y_{t_0}) + h\mathsf{C}_{\mathsf{Lip}} + \varepsilon\mathsf{C}_{\mathsf{eps}},\end{aligned}$$

where note that $Nh < 1$. Take $t_0 = 0$, $t_N = T < 1$, we obtain

$$\text{TV}\,(\pi_T, \hat{\pi}_T) \leq h\mathsf{C}_{\mathsf{Lip}} + \varepsilon\mathsf{C}_{\mathsf{eps}}.$$

□

B.6 Example: Hypersphere

Lemma 13. *On $\mathcal{S}^d$, Assumption 2 holds with the following constants.*

1. $L_t^{v,x} = \frac{12\pi(d-1)}{(1-t)}$, $L_t^{v,t} := \frac{20\pi^2 d}{1-t}$.
2. $L_t^{Div,x} := \frac{128\pi(d-1)^2}{(1-t)^2}$, $L_t^{Div,t} := \frac{256\pi^2(d-1)^2}{(1-t)^2}$.
3. $L_t^{score} := \frac{8(d-1)^2}{(1-t)^2}$.
4. $L^v = \pi$.
5. $L_R = 1$.

As the proof is technical, we defer it to Appendix D.

Now we prove Proposition 3.

Proof. [Proof of Proposition 3] We first recall the extra condition on h imposed by Lemma 5:

$$h < \min\{\frac{R}{\|b(x)\|}, \frac{1}{4L_\nabla}, \sqrt{\frac{3}{4\|b(x)\|^2 L_R(2+2L_\nabla \max\{\frac{1}{\sqrt{K_{\min}}}, 1\})}}\}.$$

Lemma 18 additionally requires $h^2 L_R (L^{\hat{v}})^2 \le \frac{1}{4}$. Under Assumption 3, since $\|\hat{v}(t,x)\| \le L^{\hat{v}} \lesssim L^v$, we know $\|b(x)\|$ is of constant order, consequently $\frac{R}{\|b(x)\|}$ is of constant order. Thus such an condition is dominated by the term $\frac{1}{L_\nabla}$, which in our case is exactly $\frac{1}{L_t^{\hat{v},x}}$. We already have the following two bounds:

$$\begin{aligned}
&\mathbb{E}[\|\text{ Grad } \log p(t,x)\| \cdot \|\tilde{v}(x,t) - v(x,t)\|] \\
\le&\varepsilon\sqrt{3L_t^{\text{score}}} + \Big(L_t^{\text{score}}(6(t-t_k)^2\Big((L_t^{\hat{v},x})^2(L_t^{\hat{v}})^2 + (L_t^{\hat{v},x})^2(L_t^v)^2 + (L_t^{v,t})^2 + (L_t^v)^2(L_t^{v,x})^2\Big)\Big)^{\frac{1}{2}}, \\
&\mathbb{E}_{p(t,x)}[|\text{ Div }(\tilde{v}(x,t) - v(x,t))|] \\
\le&3\varepsilon + (t-t_k)\Big(L_t^{\text{Div }\hat{v},x}(L_t^{\hat{v}} + L_t^v) + L_t^{\text{Div},x}L_t^v + L_t^{\text{Div},t} + 5L_R(L_t^{\hat{v}})^2 d + 2(L_t^{\hat{v},x})^2\Big).
\end{aligned}$$

Plug in constants in Lemma 13, we obtain

$$\begin{aligned}
\mathbb{E}[\|\text{ Grad } \log p(t,x)\| \cdot \|\tilde{v}(x,t) - v(x,t)\|] \lesssim&\varepsilon\sqrt{\frac{(d-1)^2}{(1-t)^2}} + (t-t_k)\Big(\frac{(d-1)^2}{(1-t)^2}(\frac{(d-1)^2}{(1-t)^2})\Big)^{\frac{1}{2}} \\
\lesssim&\varepsilon\frac{d-1}{1-t} + (t-t_k)\frac{(d-1)^2}{(1-t)^2},
\end{aligned}$$

and

$$\mathbb{E}_{p(t,x)}[|\text{ Div }(\tilde{v}(x,t) - v(x,t))|] \lesssim \varepsilon + (t-t_k)\frac{(d-1)^2}{(1-t)^2}.$$

With early stopping, we terminate the sampling algorithm (Algorithm 1) at time $T < 1$ (i.e., $t_0 = 0, t_N = T$)

$$\begin{aligned}
&\text{TV}(\pi_T, \hat{\pi}_T) = \text{TV}(X_{t_N}, Y_{t_N}) \\
\le&\text{TV}(X_{t_0}, Y_{t_0}) + \int_{t_0}^{t_N} \mathbb{E}[\|\text{ Grad } \log p(t,x)\| \cdot \|\tilde{v}(x,t) - v(x,t)\|]dt + \int_{t_0}^{t_N} \mathbb{E}[|\text{ Div }(\tilde{v}(x,t) - v(x,t))|]dt \\
\lesssim&\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}} \varepsilon\frac{d-1}{1-t} + (t-t_i)\frac{(d-1)^2}{(1-t)^2}dt + \sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}} \varepsilon + (t-t_i)\frac{(d-1)^2}{(1-t)^2}dt \\
\lesssim&(d-1)\varepsilon\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}} \frac{1}{1-t}dt + (d-1)^2\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}} \frac{t-t_i}{(1-t)^2}dt.
\end{aligned}$$

We first discuss the case of constant step size. Using

$$\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}}\frac{1}{1-t}dt=-\log(1-t_N)$$

and (using $s:=1-t$ so that $ds=-dt$)

$$\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}}\frac{t-t_i}{(1-t)^2}dt\le h\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}}\frac{1}{(1-t)^2}dt=h\sum_{i=0}^{N-1}\int_{1-t_{i+1}}^{1-t_i}\frac{1}{s^2}$$

$$=h\sum_{i=0}^{N-1}-\frac{1}{(1-t_i)^1}+\frac{1}{(1-t_{i+1})^1}\lesssim h\frac{1}{1-T},$$

we can finally bound the error as

$$\varepsilon(d-1)\log(\frac{1}{1-T})+h(d-1)^2\frac{1}{1-T}.$$

Now we have

$$\mathrm{TV}\left(X_T,Y_T\right)\lesssim\varepsilon(d-1)\log(\frac{1}{1-T})+h(d-1)^2\frac{1}{1-T}.$$

To obtain a sample up to $\varepsilon_{\text{target}}$ accuracy, we assume ε is sufficiently small. Then we need

$$h(d-1)^2\frac{1}{1-T}=\mathcal{O}(\varepsilon_{\text{target}}),$$

which means

$$h=\frac{\varepsilon_{\text{target}}(1-T)}{(d-1)^2}. \tag{11}$$

Next, we discuss a specific step size schedule that can improve the dependency on $\frac{1}{1-T}$. Denote $h_k=t_{k+1}-t_k$. We choose some constant $\alpha\in(0,\frac{1}{2})$, and $h_i=\alpha(1-t_i)$. Equivalently,

$$t_i=1-(1-\alpha)^i,\qquad 1-t_{i+1}=(1-\alpha)^{i+1}=(1-\alpha)(1-t_i),\qquad\text{and}\qquad T=t_N=1-(1-\alpha)^N. \tag{12}$$

Note that the last equation implies $N=\frac{\log(1-T)}{\log(1-\alpha)}$ Then

$$\int_{t_i}^{t_{i+1}}\frac{1-t_i}{(1-t)^2}dt=(1-t_i)(\frac{1}{1-t_{i+1}}-\frac{1}{1-t_i})=\frac{h_i(1-t_i)}{(1-t_{i+1})(1-t_i)}=\frac{\alpha}{1-\alpha},$$

$$\int_{t_i}^{t_{i+1}}\frac{t-1}{(1-t)^2}dt=-\int_{t_i}^{t_{i+1}}\frac{1}{1-t}dt=-\log(1-t_i)+\log(1-t_{i+1})=\log(1-\alpha).$$

Together,

$$\int_{t_i}^{t_{i+1}}\frac{t-t_i}{(1-t)^2}dt=\frac{\alpha}{1-\alpha}+\log(1-\alpha)\le\frac{\alpha^2}{1-\alpha},\forall\alpha\in(0,\frac{1}{2}),$$

which implies

$$\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}}\frac{t-t_i}{(1-t)^2}dt\lesssim N\frac{\alpha^2}{1-\alpha}\lesssim\frac{\alpha^2}{1-\alpha}\frac{\log(1-T)}{\log(1-\alpha)}\le2\alpha\log\frac{1}{1-T},\forall\alpha\in(0,\frac{1}{2}).$$

We want to reach

$$(d-1)^2\alpha\log\frac{1}{1-T}=\mathcal{O}(\varepsilon_{\text{target}}).$$

So we need

$$\alpha\lesssim\Big(\frac{\varepsilon_{\text{target}}}{(d-1)^2\log\frac{1}{1-T}}\Big).$$

and consequently

$$N=\frac{\log(1-T)}{\log(1-\alpha)}=\frac{\log(1-T)}{\log(1-\frac{\varepsilon_{\text{target}}}{(d-1)^2\log\frac{1}{1-T}})}=\mathcal{O}(\frac{1}{\frac{\varepsilon_{\text{target}}}{(d-1)^2\log\frac{1}{1-T}}}\log\frac{1}{1-T})=\mathcal{O}(\frac{(d-1)^2}{\varepsilon_{\text{target}}}\log^2\frac{1}{1-T})$$

□

B.7 Hadamard Manifold with $\nabla R \equiv 0$

We first state the following result proved later in Appendix E.

Proposition 14. *Let M be a Hadamard manifold that is also a symmetric space. Assume Asumption 1. We impose the following moment condition: there exists M_λ be such that*

$$\max\left\{\mathbb{E}[d(X_1,z)^2 e^{24\lambda d(X_1,z)}], \mathbb{E}[e^{24\lambda d(X_1,z)}]\right\} \le M_\lambda, \qquad \textit{where} \qquad \lambda = \max\{1,\kappa\}.$$

We choose the prior distribution to be a Riemannian Gaussian distribution centered at some $z \in M$: $p_0(x) \propto \exp\big(-d\, d(x,z)^2\big)$. We then have the following regularity results.

$$\begin{aligned}
\mathbb{E}[\|v(t,x)\|^2] &\lesssim d M_\lambda^{\frac{1}{2}}, \\
\mathbb{E}[\|\nabla v(t,x)\|] &\lesssim \frac{d^{2+6\lambda}}{1-t} L_R M_\lambda^{\frac{1}{2}}, & \mathbb{E}[\|\nabla v(t,x)\|^2] &\lesssim \frac{d^{3+12\lambda}}{(1-t)^2} L_R^2 M_\lambda^{\frac{1}{2}}, \\
\mathbb{E}[|\frac{d}{dt} v(t,x)|] &\lesssim \frac{d^{2+6\lambda}}{1-t} L_R M_\lambda^{\frac{1}{2}}, & \mathbb{E}[|\frac{d}{dt} v(t,x)|^2] &\lesssim \frac{d^{3+12\lambda}}{(1-t)^2} L_R^2 M_\lambda^{\frac{1}{2}}, \\
\mathbb{E}[\|\textit{ Grad }_x \textit{ Div } v(t,x)\|] &\lesssim \frac{d^{3+12\lambda}}{(1-t)^2} L_R^3 M_\lambda^{\frac{1}{2}}, & \mathbb{E}[\|\textit{ Grad }_x \textit{ Div } v(t,x)\|^2] &\lesssim \frac{d^{5+24\lambda}}{(1-t)^4} L_R^6 M_\lambda, \\
\mathbb{E}[|\frac{d}{dt} \textit{ Div } v(t,x)|] &\lesssim \frac{d^{3+12\lambda}}{(1-t)^2} L_R^3 M_\lambda^{\frac{1}{2}}, & \mathbb{E}[\|\nabla v(t,x)\|_{\mathrm{HS}}^2] &\lesssim \frac{d^{3+12\lambda}}{(1-t)^2} L_R^2 M_\lambda^{\frac{1}{2}},
\end{aligned}$$

where $\lambda = \max\{1,\kappa\}$.

Proof. [Proof of Proposition 2]

Using Proposition 14 and Proposition 7, we have

$$\begin{aligned}
&\mathrm{TV}\left(\pi_T, \hat{\pi}_T\right) = \mathrm{TV}\left(X_{t_N}, Y_{t_N}\right) \\
&\lesssim \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} (t-t_i)\Big(\sqrt{\frac{d^{5+12\lambda}}{(1-t)^2} L_R^2 M_\lambda M_\lambda \frac{d^2}{(1-t)^2} L_R^2 d^{12\lambda}} + \frac{d^{4+12\lambda}}{(1-t)^2} L_R^3 M_\lambda^{\frac{1}{2}}\Big) dt \\
&\quad + \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \Big(\varepsilon \sqrt{M_\lambda \frac{d^2}{(1-t)^2} L_R^2 d^{12\lambda}} + \varepsilon\Big) dt \\
&\lesssim \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} (t-t_i) \frac{d^{4+12\lambda}}{(1-t)^2} L_R^3 M_\lambda dt + \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \Big(\frac{d^{6\lambda+1}}{1-t} L_R\Big) M_\lambda^{\frac{1}{2}} \varepsilon dt.
\end{aligned}$$

We apply the same step size schedule as before: $h_k = t_{k+1} - t_k$, constant $\alpha \in (0, \frac{1}{2})$, and $h_i = \alpha(1-t_i)$. Equivalently,

$$t_i = 1-(1-\alpha)^i, \qquad 1-t_{i+1} = (1-\alpha)^{i+1} = (1-\alpha)(1-t_i), \qquad \text{and} \qquad T = t_N = 1-(1-\alpha)^N.$$

We emphasize that, the difference between Hadamard case and compact manifold case is that, the upper bound on step size imposed by Lemma 5 would be affected by $K_{\min}$. More specifically,

$$h < \min\{\frac{R}{B}, \frac{1}{4L_\nabla}, \sqrt{\frac{3}{4\|b(x)\|^2 L_R (2\frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}} + 4\frac{\cosh(\sqrt{-K_{\min}})-1}{-K_{\min}} L_\nabla)}}\},$$

and we can write this upper bound as

$$h \lesssim \frac{1}{L_\nabla \sqrt{\frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}}}} \lesssim \frac{1-t}{\sqrt{\frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}}} d^{2+6\lambda} L_R M_\lambda^{\frac{1}{2}}}$$

note that when $K_{\min}$ is not of constant order, h must be exponentially small w.r.t. $K_{\min}$.

Following the step size schedule, Thus to reach $\varepsilon_{\text{target}}$ accuracy, we need to reach

$$d^{4+12\lambda} L_R^3 M_\lambda \alpha \log \frac{1}{1-T} = \mathcal{O}(\varepsilon_{\text{target}}).$$

So we need

$$\alpha \lesssim \Big(\frac{\varepsilon_{\text{target}}}{\sqrt{\frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}}} d^{4+12\lambda} L_R^3 M_\lambda \log \frac{1}{1-T}} \Big),$$

where the extra factor $\sqrt{\frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}}}$ comes from Lemma 5. Consequently,

$$\begin{aligned} N = \frac{\log(1-T)}{\log(1-\alpha)} &= \frac{\log(1-T)}{\log(1 - \frac{\varepsilon_{\text{target}}}{\sqrt{\frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}}} d^{4+12\lambda} L_R^3 M_\lambda \log \frac{1}{1-T}})} \\ &= \mathcal{O}(\frac{1}{\frac{\varepsilon_{\text{target}}}{\sqrt{\frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}}} d^{4+12\lambda} L_R^3 M_\lambda \log \frac{1}{1-T}}} \log \frac{1}{1-T}) \\ &= \mathcal{O}(\frac{\sqrt{\frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}}} d^{4+12\lambda} L_R^3 M_\lambda}{\varepsilon_{\text{target}}} \log^2 \frac{1}{1-T}) \end{aligned}$$

We verify that such step size design satisfy the requirements of Lemma 5 and Lemma 18. For small $\varepsilon_{\text{target}}$, we have

$$\begin{aligned} h_i = \alpha(1-t_i) &\lesssim \frac{\varepsilon_{\text{target}}(1-t_i)}{\sqrt{\frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}}} d^{4+12\lambda} L_R^3 M_\lambda \log \frac{1}{1-T}} \\ &\leq \frac{1-t_i}{\sqrt{\frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}}} d^{2+6\lambda} L_R M_\lambda^{\frac{1}{2}}}. \end{aligned}$$

$$h_i^2 L_R (L_{t_i}^{\hat{v}})^2 \lesssim \alpha^2 (1-t_i)^2 L_R d \leq \frac{1}{4}.$$

Therefore, we conclude the iteration complexity as

$$N = \mathcal{O}(\frac{\sqrt{\frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}}} d^{4+12\lambda} L_R^3 M_\lambda}{\varepsilon_{\text{target}}} \log^2 \frac{1}{1-T}).$$

We remark that on $\mathrm{SPD}\,(n)$, since $K_{\min} = -\frac{1}{2}$ [9], so $\lambda = \max\{1, \kappa\} = 1$. In the meanwhile, the upper bound on h required by Lemma 5 can be reduced to be of order

$$\min\{\frac{1}{L_\nabla}, \sqrt{\frac{1}{\|b(x)\|^2 L_R L_\nabla}}\} = \frac{1-t}{d^{2+6\lambda} L_R M_\lambda^{\frac{1}{2}}},$$

which is satisfied by our design $h_i = \alpha(1-t_i)$. Then we have

$$N = \mathcal{O}(\frac{d^{16} L_R^3 M_\lambda}{\varepsilon_{\text{target}}} \log^2 \frac{1}{1-T})$$

□

C Auxiliary Results for Proof of Main Theorems

C.1 Jacobi Equation

Lemma 15 (Solution for Jacobi equation)**.** *Let M be of constant sectional curvature c. Let J be a normal Jacobi field along γ, with initial condition $J(0) = v^\perp$, $J'(0) = 0$. Then we have*

$$J(t) = s_c^{(2)}(t)\|v^\perp\| E(t) := \begin{cases} \|v^\perp\| E(t), & \text{if } c = 0, \\ \|v^\perp\| \cos(\sqrt{c}t) E(t), & \text{if } c > 0, \\ \|v^\perp\| \cosh(\sqrt{-c}t) E(t), & \text{if } c < 0. \end{cases}$$

where E is a parallel normal unit vector field with $E(0) = \frac{v^\perp}{\|v^\perp\|}$.

Proof. [Proof of Lemma 15] Similar to the proof of [21, Proposition 10.12], the solution is of the form $J(t) = f(t)E(t)$ where $E(t)$ is a parallel unit normal vector field along γ. Since the curvature is constant and J is a normal Jacobi field, the Jacobi equation reduces to $D_t^2 J + cJ = 0$, thus we only need to solve for $f''(t) + cf(t) = 0$, where $f(t) \in \mathbb{R}, \forall t$.

- When $c = 0$, we obtain $f''(t) = 0$, hence (to satisfy initial condition) $f(t) = \|v^\perp\|$.
- When $c > 0$, we obtain $f(t) = \|v^\perp\| \cos(\sqrt{c}t)$.
- When $c < 0$, we obtain $f(t) = \|v^\perp\| \frac{e^{\sqrt{-c}t} + e^{-\sqrt{-c}t}}{2} = \|v^\perp\| \cosh(\sqrt{c}t)$.

Note that $E(0) = \frac{v^\perp}{\|v^\perp\|}$. Therefore,

$$J(t) = \begin{cases} \|v^\perp\| E(t), & \text{if } c = 0, \\ \|v^\perp\| \cos(\sqrt{c}t) E(t), & \text{if } c > 0, \\ \|v^\perp\| \cosh(\sqrt{-c}t) E(t), & \text{if } c < 0. \end{cases}$$

□

Lemma 16. *Let $J_v(t)$ be a Jacobi field along $\gamma(t) = Exp_x(thb(x))$, with $J_v(0) = v$, $J_v'(0) = \omega$. Define $Y(t) = P_{\gamma(t)}^{\gamma(0)} J_v(t)$. Up to the first conjugate point of x, we have*

$$\|Y(t)\| \le \|v^\parallel\| + \|\omega^\parallel\| t + s_{K_{\min}}(t) \|\omega^\perp\| + \frac{s_{K_{\min}}(t)}{s_{K_{\min}}(t_0)} \|v^\perp\|.$$

Take $h < \frac{R}{\|b(x)\|}$, we have

$$\int_0^1 \|Y(t)\| dt \le \begin{cases} \|v^\parallel\| + \|\omega^\parallel\| + \frac{\|\omega^\perp\|}{\sqrt{K_{\min}}} + \|v^\perp\|, & \text{if } K_{\min} > 0, \\ \|v^\parallel\| + \|\omega^\parallel\| + \frac{\cosh(\sqrt{-K_{\min}}) - 1}{-K_{\min}} \|\omega^\perp\| + \frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}} \|v^\perp\|, & \text{if } K_{\min} < 0, \\ \|v^\parallel\| + \frac{1}{2}\|\omega^\parallel\| + \frac{1}{2}\|\omega^\perp\| + \|v^\perp\|, & \text{if } K_{\min} = 0. \end{cases}$$

Proof. [Proof of Lemma 16] Using isometric property of parallel transport, $\|Y(t)\| = \|J_v(t)\|$. Decompose $v = v^\parallel + v^\perp$, where $v^\parallel$ is the component in $\gamma'(0)$ direction, and $\langle v^\perp, \gamma'(0) \rangle = 0$. Similarly, we decompose $\omega = \omega^\parallel + \omega^\perp$.

Define $J_v =: J^{(0)} + J^{(1)} + J^{(2)}$, where

$$\begin{aligned} J^{(0)}(0) = v^\parallel, \quad & D_t J^{(0)}(0) = \omega^\parallel; \\ J^{(1)}(0) = 0, \quad & D_t J^{(1)}(0) = \omega^\perp; \\ J^{(2)}(0) = v^\perp, \quad & D_t J^{(2)}(0) = 0. \end{aligned}$$

Note that both $J^{(1)}(0)$, $D_t J^{(1)}(0)$ are orthogonla to $\gamma'(0)$, so $J^{(1)}(t)$ is a normal Jacobi field.

We know $J^{(0)}$ is a tangential Jacobi field, so it has the form

$$J^{(0)}(t) = (a + bt)\gamma'(t).$$

Plug in the initial value, we get (note that $D_t \gamma'(0) = 0$)

$$a\gamma'(0) = v^\parallel, \quad b\gamma'(0) = \omega^\parallel.$$

Hence $|a| = \frac{\|v^\parallel\|}{\|\gamma'(0)\|}$, and $|b| = \frac{\|\omega^\parallel\|}{\|\gamma'(0)\|}$.

By definition, $\|\gamma'(0)\| = \|hb(x)\| = \|\gamma'(t)\|$. So we get

$$\|J^{(0)}(t)\| \le \|v^\parallel\| + \|\omega^\parallel\| t.$$

If all sectional curvatures of M are bounded below by a constant $K_{\min}$, Jacobi field comparison theorem yield

$$\|J^{(1)}(t)\| \le s_K(t) \|D_t J^{(1)}(0)\| = s_{K_{\min}}(t) \|\omega^\perp\|.$$

Now consider the term $J^2(0)$. By [4, Theorem 1.34] applied with Jacobi field formula given in Lemma 15, we obtain $\|J^{(2)}(t)\| \leq \|\tilde{J}(t)\|$. We remark that the focal point free condition is saying $J^{(2)}(t) \neq 0$. As long as Cut locus is not reached, the geodesic is minimizing. Hence it suffices to guarantee that the geodesic γ satisfies $\|\gamma'(0)\| < \text{Inj}\,(M)$.

Finally, recall $\|Y(t)\| = \|J_v(t)\|$.

$$\begin{aligned}\|Y(t)\| =&\|J_v(t)\| = \|J^{(0)}(t) + J^{(1)}(t) + J^{(2)}(t)\| \\ \leq&\|v^{\|}\| + \|\omega^{\|}\|t + |s_{K_{\min}}(t)|\|\omega^{\perp}\| + |s^{(2)}_{K_{\min}}(t)|\|v^{\perp}\|.\end{aligned}$$

We split into cases. The first case is $K_{\min} > 0$. In this case we denote $R = \text{Inj}\,(M)$ and

$$\int_0^1 |s_{K_{\min}}(t)|dt = \frac{1}{\sqrt{K_{\min}}}\int_0^1 |\sin(t\sqrt{K_{\min}})|dt \leq \frac{1}{\sqrt{K_{\min}}}.$$

Also,

$$\int_0^1 |s^{(2)}_{K_{\min}}(t)|dt = \int_0^1 |\cos(\sqrt{K_{\min}}t)|dt \leq 1.$$

To summarize, when $h < \frac{R}{\|b(x)\|}$,

$$\int_0^1 \|Y(t)\|dt \leq \|v^{\|}\| + \|\omega^{\|}\| + \frac{\|\omega^{\perp}\|}{\sqrt{K_{\min}}} + \|v^{\perp}\|.$$

The second case is $K_{\min} < 0$.

Then $s_{K_{\min}}(t) = \frac{1}{\sqrt{-K_{\min}}}\sinh(\sqrt{-K_{\min}}t)$. Consider $0 < t \leq h < 1$.

$$\int_0^1 s_{K_{\min}}(t)dt = \int_0^1 \frac{1}{\sqrt{-K_{\min}}}\sinh t\sqrt{-K_{\min}}dt = \frac{1}{-K_{\min}}(\cosh(\sqrt{-K_{\min}}) - 1).$$

For the second integral,

$$\int_0^1 \cosh(\sqrt{-K_{\min}}t)dt = \frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}}.$$

To summarize, when $h < \frac{R}{\|b(x)\|}$,

$$\int_0^1 \|Y(t)\|dt \leq \|v^{\|}\| + \|\omega^{\|}\| + \frac{\cosh(\sqrt{-K_{\min}}) - 1}{-K_{\min}}\|\omega^{\perp}\| + \|v^{\perp}\|\frac{\sinh(\sqrt{-K_{\min}})}{\sqrt{-K_{\min}}}.$$

The third case is $K_{\min} = 0$. In this case $s_{K_{\min}}(t) = t$.

$$\begin{aligned}\|Y(t)\| =&\|J_v(t)\| = \|J^{(0)}(t) + J^{(1)}(t) + J^{(2)}(t)\| \\ \leq&\|v^{\|}\| + \|\omega^{\|}\|t + \|\omega^{\perp}\|t + \|v^{\perp}\|.\end{aligned}$$

Then we have

$$\int_0^1 \|Y(t)\|dt \leq \|v^{\|}\| + \frac{1}{2}\|\omega^{\|}\| + \frac{1}{2}\|\omega^{\perp}\| + \|v^{\perp}\|.$$

□

C.2 Divergence Term

Lemma 17. *Let c denote the geodesic with $c(0) = z$, $c(h) = x$, and let $\Lambda(s,t)$ be a smooth variation satisfying $\Lambda(0,t) = c(t)$ for $0 \leq t \leq h$. Let V be a smooth vector field. Parallel transport along the variation is taken along $t \mapsto \Lambda(s,t)$, and D_s, D_t denote the induced covariant derivatives. Then*

$$\begin{aligned}D_s\left(P^{\Lambda(s,h)}_{\Lambda(s,0)}V(\Lambda(s,0))\right)\Big|_{s=0} =& P^{c(h)}_{c(0)}\nabla_{\partial_s\Lambda(0,0)}V(z) \\ &- P^{c(h)}_{c(0)}\int_0^h P^{c(0)}_{c(\tau)}R(\partial_s\Lambda(0,\tau),\partial_t\Lambda(0,\tau))P^{c(\tau)}_{c(0)}V(z)\,d\tau.\end{aligned}$$

Proof. [Proof of Lemma 17] By construction, it holds that

$$\Lambda(0,0) = z,\ \Lambda(0,h) = x.$$

and the parallel transport $P_z^x = P_{\Lambda(0,0)}^{\Lambda(0,h)}$ is along c. Define $W(s,t) = P_{\Lambda(s,0)}^{\Lambda(s,t)} V(\Lambda(s,0))$, where we perform parallel transport along curve $t \mapsto \Lambda(s,t)$ which might not be a geodesic. By definition of parallel transport, $D_t W(s,t) = 0,\ \forall s$.

We know

$$-D_t D_s W(s,t) = D_s D_t W(s,t) - D_t D_s W(s,t) = R(\partial_s \Lambda(s,t), \partial_t \Lambda(s,t)) W(s,t).$$

Hence evaluating at $s = 0$, we obtain

$$D_t D_s W(0,t) = -R(\partial_s \Lambda(0,t), \partial_t \Lambda(0,t)) W(0,t).$$

We perform parallel transport $P_{c(t)}^{c(0)}$ on both sides of the equation, and by Lemma 19 we have

$$\frac{d}{dt} P_{c(t)}^{c(0)} D_s W(0,t) = P_{c(t)}^{c(0)} D_t D_s W(0,t) = -P_{c(t)}^{c(0)} R(\partial_s \Lambda(0,t), \partial_t \Lambda(0,t)) W(0,t).$$

Observe that both sides of the equation are time dependent vector fields in $T_{c(0)}M = T_z M$. Hence we can perform integration

$$\int_0^h \frac{d}{d\tau} P_{c(\tau)}^{c(0)} D_s W(0,\tau) d\tau = -\int_0^h P_{c(\tau)}^{c(0)} R(\partial_s \Lambda(0,\tau), \partial_t \Lambda(0,\tau)) W(0,\tau) d\tau.$$

Hence

$$P_{c(h)}^{c(0)} D_s W(0,h) = D_s W(0,0) - \int_0^h P_{c(\tau)}^{c(0)} R(\partial_s \Lambda(0,\tau), \partial_t \Lambda(0,\tau)) W(0,\tau) d\tau.$$

Note that $D_s W(0,0) = \nabla_{\partial_s \Lambda(0,0)} V(z)$. Perform parallel transport $P_{c(0)}^{c(h)}$ on both sides of the equation, we have

$$D_s W(0,h) = P_{c(0)}^{c(h)} \nabla_{\partial_s \Lambda(0,0)} V(z) - P_{c(0)}^{c(h)} \int_0^h P_{c(\tau)}^{c(0)} R(\partial_s \Lambda(0,\tau), \partial_t \Lambda(0,\tau)) W(0,\tau) d\tau.$$

Recall $W(s,t) = P_{\Lambda(s,0)}^{\Lambda(s,t)} V(\Lambda(s,0))$.

$$D_s W(0,h) = P_{c(0)}^{c(h)} \nabla_{\partial_s \Lambda(0,0)} V(z) - P_{c(0)}^{c(h)} \int_0^h P_{c(\tau)}^{c(0)} R(\partial_s \Lambda(0,\tau), \partial_t \Lambda(0,\tau)) P_{c(0)}^{c(\tau)} V(z) d\tau.$$

□

Notice that $\tilde{v}$ is defined through $\hat{v}$ and the inverse of F. We need to control Div $\tilde{v}(t,x)$ in our analysis, by writing it as some expression involving Div $\hat{v}$.

Lemma 18. *Under Assumptions 1 and 3, take $B = L_{t_k}^{\hat{v}}$ and $L_\nabla = L_{t_k}^{\hat{v},x}$ in Lemma 5. Let $h = t - t_k \geq 0$ satisfy its step-size requirements and $hL_\nabla \leq 1/4$, $h^2 L_R B^2 \leq 1/4$. Then, for $z = F_{t_k,h}^{-1}(x)$,*

$$|\textit{Div}\,(\tilde{v}(t,x) - v(t,x))| \leq |\textit{Div}\ \hat{v}(t_k,z) - \textit{Div}\ v(t,x)| + h\left(2L_\nabla^2 + 5dL_R \|\hat{v}(t_k,z)\|^2\right).$$

Proof. The case $h = 0$ follows from $\tilde{v}(t_k,x) = \hat{v}(t_k,x)$. For $h > 0$, we follow the setting in Lemma 17 with V replaced by $\hat{v}(t_k,\cdot)$ and the time interval shifted to $[t_k,t]$. Set $c(\tau) = \mathrm{Exp}_z((\tau - t_k)\hat{v}(t_k,z))$ and let $\{E_i(\tau)\}_{i=1}^d$ be a parallel orthonormal frame along c. For each i, define

$$\Lambda_i(s,t_k) = \mathrm{Exp}_z\left(s(dF_{t_k,h})_z^{-1} E_i(t)\right), \qquad \Lambda_i(s,\tau) = F_{t_k,\tau-t_k}(\Lambda_i(s,t_k)).$$

Then $\Lambda_i(0,\tau) = c(\tau)$, $\partial_\tau c(\tau) = P_z^{c(\tau)} \hat{v}(t_k,z)$, and

$$\partial_s \Lambda_i(0,\tau) = (dF_{t_k,\tau-t_k})_z (dF_{t_k,h})_z^{-1} E_i(t), \qquad \partial_s \Lambda_i(0,t) = E_i(t).$$

The identity $F_{t_k,h}^{-1}(\Lambda_i(s,t)) = \Lambda_i(s,t_k)$ and the definition of the interpolation give

$$P_{\Lambda_i(s,t_k)}^{\Lambda_i(s,t)} \hat{v}(t_k, \Lambda_i(s,t_k)) = \tilde{v}(t, \Lambda_i(s,t)).$$

Thus Lemma 17 differentiates the required vector field in direction $E_i(t)$. By definition of Riemannian divergence and the isometric property of parallel transport,

$$\begin{aligned}
\text{Div } \tilde{v}(t,x) &= \sum_{i=1}^{d} \langle \nabla_{E_i(t)} \tilde{v}(t,x), E_i(t) \rangle \\
&= \sum_{i=1}^{d} \langle \nabla_{\partial_s \Lambda_i(0,t_k)} \hat{v}(t_k,z), E_i(t_k) \rangle \\
&\quad - \sum_{i=1}^{d} \int_{t_k}^{t} \langle R(\partial_s \Lambda_i(0,\tau), \partial_\tau c(\tau)) \partial_\tau c(\tau), E_i(\tau) \rangle \, d\tau.
\end{aligned}$$

Subtracting Div $\hat{v}(t_k,z) = \sum_{i=1}^{d} \langle \nabla_{E_i(t_k)} \hat{v}(t_k,z), E_i(t_k) \rangle$ and using Assumption 1, we obtain

$$\begin{aligned}
|\text{Div } \tilde{v}(t,x) - \text{Div } \hat{v}(t_k,z)| &\le \left| \sum_{i=1}^{d} \langle \nabla_{\partial_s \Lambda_i(0,t_k) - E_i(t_k)} \hat{v}(t_k,z), E_i(t_k) \rangle \right| \\
&\quad + L_R \|\hat{v}(t_k,z)\|^2 \sum_{i=1}^{d} \int_{t_k}^{t} \|\partial_s \Lambda_i(0,\tau)\| \, d\tau. \qquad (13)
\end{aligned}$$

The variation is through geodesics, so

$$D_\tau^2 \partial_s \Lambda_i(0,\tau) = -R(\partial_s \Lambda_i(0,\tau), \partial_\tau c(\tau)) \partial_\tau c(\tau).$$

At $\tau = t_k$, $D_\tau \partial_s \Lambda_i(0,t_k) = \nabla_{\partial_s \Lambda_i(0,t_k)} \hat{v}(t_k,z)$. Applying Lemma 19 and integrating twice gives

$$\begin{aligned}
P_{c(\tau)}^{z} \partial_s \Lambda_i(0,\tau) &= \partial_s \Lambda_i(0,t_k) + (\tau - t_k) \nabla_{\partial_s \Lambda_i(0,t_k)} \hat{v}(t_k,z) \\
&\quad - \int_{t_k}^{\tau} (\tau - r) P_{c(r)}^{z} R(\partial_s \Lambda_i(0,r), \partial_r c(r)) \partial_r c(r) \, dr. \qquad (14)
\end{aligned}$$

Taking norms and then the supremum over $\tau \in [t_k, t]$, and using the step-size restrictions, yields

$$\sup_{t_k \le \tau \le t} \|\partial_s \Lambda_i(0,\tau)\| \le \frac{1 + hL_\nabla}{1 - h^2 L_R \|\hat{v}(t_k,z)\|^2/2} \|\partial_s \Lambda_i(0,t_k)\| \le 2 \|\partial_s \Lambda_i(0,t_k)\|. \qquad (15)$$

Evaluating (14) at $\tau = t$ and using (15), we obtain

$$1 = \|E_i(t_k)\| \ge \left(1 - hL_\nabla - h^2 L_R \|\hat{v}(t_k,z)\|^2\right) \|\partial_s \Lambda_i(0,t_k)\| \ge \tfrac{1}{2} \|\partial_s \Lambda_i(0,t_k)\|.$$

Consequently, $\|\partial_s \Lambda_i(0,t_k)\| \le 2$ and $\sup_{t_k \le \tau \le t} \|\partial_s \Lambda_i(0,\tau)\| \le 4$. Replacing $E_i(t)$ by any $u \in T_x M$ in the variation gives $\|u\| \ge \frac{1}{2} \|(dF_{t_k,h})_z^{-1} u\|$, hence $\|(dF_{t_k,h})_z^{-1}\|_{\text{op}} \le 2$. Let $(\nabla \hat{v}(t_k,z))^*$ denote the g_z-adjoint. Since $\partial_s \Lambda_i(0,t_k) = (dF_{t_k,h})_z^{-1} P_z^x E_i(t_k)$, Cauchy–Schwarz gives

$$\begin{aligned}
&\left| \sum_{i=1}^{d} \langle \nabla_{\nabla_{\partial_s \Lambda_i(0,t_k)} \hat{v}(t_k,z)} \hat{v}(t_k,z), E_i(t_k) \rangle \right| \\
&= \left| \sum_{i=1}^{d} \langle \nabla_{\partial_s \Lambda_i(0,t_k)} \hat{v}(t_k,z), (\nabla \hat{v}(t_k,z))^* E_i(t_k) \rangle \right| \\
&\le \left\| \nabla \hat{v}(t_k,z) \circ (dF_{t_k,h})_z^{-1} \circ P_z^x \right\|_{\text{HS}} \|\nabla \hat{v}(t_k,z)\|_{\text{HS}} \\
&\le \|(dF_{t_k,h})_z^{-1}\|_{\text{op}} \|\nabla \hat{v}(t_k,z)\|_{\text{HS}}^2 \le 2L_\nabla^2,
\end{aligned}$$

where the last inequality uses the tensor norm bound in Assumption 3. Using (14) at $\tau = t$ to express $\partial_s \Lambda_i(0,t_k) - E_i(t_k)$, we therefore obtain

$$\left| \sum_{i=1}^{d} \langle \nabla_{\partial_s \Lambda_i(0,t_k) - E_i(t_k)} \hat{v}(t_k,z), E_i(t_k) \rangle \right|$$

$$\leq h\left|\sum_{i=1}^{d}\langle\nabla_{\nabla_{\partial_s\Lambda_i(0,t_k)}\hat{v}(t_k,z)}\hat{v}(t_k,z),E_i(t_k)\rangle\right|+L_\nabla L_R\|\hat{v}(t_k,z)\|^2\sum_{i=1}^{d}\int_{t_k}^{t}(t-r)\|\partial_s\Lambda_i(0,r)\|\,dr$$
$$\leq 2hL_\nabla^2+2dh^2L_\nabla L_R\|\hat{v}(t_k,z)\|^2.$$

The second term in (13) is at most $4dhL_R\|\hat{v}(t_k,z)\|^2$. Substituting these bounds into (13) gives

$$\begin{aligned}|\,\text{Div}\;\tilde{v}(t,x)-\text{Div}\;\hat{v}(t_k,z)|&\leq 2hL_\nabla^2+dh(4+2hL_\nabla)L_R\|\hat{v}(t_k,z)\|^2\\&\leq h\left(2L_\nabla^2+5dL_R\|\hat{v}(t_k,z)\|^2\right),\end{aligned}$$

where we used $hL_\nabla\leq 1/4$. The asserted bound follows by triangle inequality.

□

Lemma 19. *We have that*

$$\frac{d}{dt}P_{c(t)}^{c(0)}Y(t)=P_{c(t)}^{c(0)}D_tY(t).$$

Proof. [Proof of Lemma 19] Denote the geodesic along t direction as $c(t)$. Let $v\in T_{c(0)}M$ be arbitrary, and define $Z(t)=P_{c(0)}^{c(t)}v$.

$\frac{d}{dt}\langle P_{c(t)}^{c(0)}Y(t),v\rangle=\frac{d}{dt}\langle Y(t),P_{c(0)}^{c(t)}v\rangle=\langle D_tY(t),P_{c(0)}^{c(t)}v\rangle+0=\langle D_tY(t),P_{c(0)}^{c(t)}v\rangle=\langle P_{c(t)}^{c(0)}D_tY(t),v\rangle.$
On the other hand,

$$\frac{d}{dt}\langle P_{c(t)}^{c(0)}Y(t),v\rangle=\langle\frac{d}{dt}P_{c(t)}^{c(0)}Y(t),v\rangle.$$

Since the above holds for any v, we have

$$\frac{d}{dt}P_{c(t)}^{c(0)}Y(t)=P_{c(t)}^{c(0)}D_tY(t).$$

□

D Hypersphere Regularity Results

Recall that

$$v(t,x)=\frac{1}{1-t}\int_M\text{Log}_x(x_1)\,p_t(x_1|x)\,dV_g(x_1).$$

To establish regularity of v, it is natural to study the formula for conditional density, $p_t(x_1|x)$. In a Euclidean space, with Gaussian distribution as prior p_0, we have that $p_t(x_1\mid x_t=x)\propto p_1(x_1)\exp(-\frac{\|tx_1-x\|^2}{2(1-t)^2})$. This is a standard result for flow matching, see for example [13] and [44]. But on a Riemannian manifold, the curvature would introduce an extra term due to the change of variable formula, and the existence of cut points would introduce an extra indicator function. We provide the formula for conditional density on the hypersphere $\mathcal{S}^d$ with uniform distribution as prior, and geodesic interpolation. See Lemma 20 below.

Lemma 20 (Conditional density on S^d for geodesic interpolation)**.** *Let S^d be the unit sphere with round metric, $d\geq 2$. Let $X_1\sim p_1$ be the data distribution with smooth densities $p_1>0$ w.r.t. dV_g, and $X_0\sim p_0$ being uniform distribution independent of X_1. Consider geodesic interpolation with minimizing geodesic. Fix $t\in[0,1)$ and $x,x_1\in S^d$. Write $r=d(x,x_1)$. Denote*

$$\begin{aligned}J_t(x\mid x_1)&=\frac{1}{1-t}\Big(\frac{\sin\big(r/(1-t)\big)}{\sin r}\Big)^{d-1}\mathbf{1}_{\{d(x,x_1)<(1-t)\pi\}}\\&=\begin{cases}\frac{1}{1-t}\Big(\frac{\sin\big(r/(1-t)\big)}{\sin r}\Big)^{d-1}, & \text{if } r<(1-t)\pi,\\0, & \text{if } r\geq(1-t)\pi.\end{cases}\end{aligned}$$

Then the conditional density of X_1 given $X_t=x$ is

$$p_t(x_1\mid x)=\frac{p_1(x_1)\,J_t(x\mid x_1)}{\displaystyle\int_{S^d}p_1(z)\,J_t(x\mid z)dV_g(z)}.$$

Proof. The proof strategy is as follows. We first write out the joint distribution of X_0, X_1, and then use the change of variable formula to obtain joint distribution of X_t, X_1.

We apply the change of variable formula. We need a diffeomorphism between (X_0, X_1) and (X_t, X_1). Fix $t \in (0,1)$ and $x_1 \in S^d$. Use polar coordinates at x_1: every $y \in S^d \setminus \{-x_1\}$ can be written uniquely as $y = \mathrm{Exp}_{x_1}(r\omega)$ for some $r \in (0,\pi), \omega \in S^{d-1}$. And we know the Riemannian volume element is

$$dV_g(y) = (\sin r)^{d-1} dr\, d\omega.$$

In particular, for any x_0, we set $r_0 = d(x_0, x_1) \in (0,\pi)$ and $\omega = \frac{\mathrm{Log}_{x_1}(x_0)}{\|\mathrm{Log}_{x_1}(x_0)\|}$. Then we can write

$$x_0 = \mathrm{Exp}_{x_1}(r_0\omega), \qquad x_1 = \mathrm{Exp}_{x_1}(0 \cdot \omega).$$

and

$$X_t = \mathrm{Exp}_{x_0}\big(t\,\mathrm{Log}_{x_0}(x_1)\big) = \mathrm{Exp}_{x_1}\big((1-t)r_0\,\omega\big).$$

Thus we can define the desired diffeomorphism as $F_t : (r_0, \omega) \mapsto (r, \omega)$ satisfying

$$F_{t,x_1}(r_0,\omega) = ((1-t)r_0, \omega), \qquad F^{-1}_{t,x_1}(r,\omega) = (\frac{r}{1-t}, \omega).$$

But however, note that we have to restrict $r < (1-t)\pi$, otherwise $\frac{r}{1-t} \notin (0,\pi)$, consequently $F^{-1}_{t,x_1}(r,\omega)$ is no longer under the polar coordinate.

Now recall the change of variable formula. We should have

$$\int_{S^d} p_0(F^{-1}_{t,x_1}(x))|\det dF^{-1}_{t,x_1}| dV_g(x) = \int_{S^d} p_0(x_0) dV_g(x_0).$$

Written in polar coordinates: the volume element at x_0 and x are

$$dV_g(x_0) = (\sin r_0)^{d-1} dr_0\, d\omega, \qquad dV_g(x) = (\sin r)^{d-1} dr\, d\omega.$$

Define a function $J_t(x \mid x_1)$ to satisfy

$$dV_g(x_0) = J_t(x \mid x_1)\, dV_g(x).$$

Using $r_0 = r/(1-t)$ (hence $dr_0 = dr/(1-t)$), we get

$$(\sin r_0)^{d-1} dr_0\, d\omega = J_t(x \mid x_1)(\sin r)^{d-1} dr\, d\omega$$
$$(\sin(r/(1-t)))^{d-1} \frac{dr}{1-t} d\omega = J_t(x \mid x_1)(\sin r)^{d-1} dr\, d\omega,$$

Now we derive the density. Note that by our construction of geodesic interpolation, given any x_0 and t, the resulting x must satisfy $d(x_1, x) = (1-t)d(x_1, x_0) \le (1-t)\pi$. Hence

$$J_t(x \mid x_1) = \frac{1}{1-t}\Big(\frac{\sin\big(r/(1-t)\big)}{\sin r}\Big)^{d-1}, \qquad r = d(x, x_1),$$

for all $r < (1-t)\pi$, and $J_t(x \mid x_1) = 0$ otherwise. Thus, we can equivalently write

$$J_t(x \mid x_1) = \frac{1}{1-t}\Big(\frac{\sin\big(r/(1-t)\big)}{\sin r}\Big)^{d-1} \mathbf{1}_{\{d(x,x_1) < (1-t)\pi\}}.$$

The joint density of (X_0, X_1) is

$$p(x_0, x_1) = p_0(x_0)p_1(x_1)$$

with respect to $dV_g(x_0)\, dV_g(x_1)$. By the change of variable formula,

$$p_t(x, x_1) = p_0(F^{-1}_{t,x_1}(x))p_1(x_1)J_t(x \mid x_1).$$

Finally, since $p_t(x, x_1) = p_t(x_1 \mid x)p_t(x) = p_t(x_1 \mid x) \int_{S^d} p_t(x, z)\, dV_g(z)$, the conditional density of X_1 given $X_t = x$ is

$$p_t(x_1 \mid x) = \frac{p_t(x, x_1)}{\int_{S^d} p_t(x, z)\, dV_g(z)},$$

which gives the formula stated in the lemma.

□

Following Lemma 20, we can derive the following formula for interpolated density.

Lemma 21. *We can write the interpolated density as*

$$p_t(x) = \int_{S^d} p_t(x, x_1)\, dV_g(x_1) = \frac{1}{\mathrm{Vol}\,(S^d)} \int_{S^d} p_1(x_1) J_t(x \mid x_1) dV_g(x_1).$$

Proof. Following Lemma 20, we have

$$p_t(x, x_1) = p_0(F^{-1}_{t,x_1}(x)) p_1(x_1) J_t(x \mid x_1).$$

Consequently, we have (note that p_0 is uniform distribution)

$$p_t(x) = \int_{S^d} p_t(x, x_1)\, dV_g(x_1) = \frac{1}{\mathrm{Vol}\,(S^d)} \int_{S^d} p_1(x_1) J_t(x \mid x_1) dV_g(x_1).$$

□

Before justifying the regularity of v, we first need to show that its smoothness is not destroyed by the indicator function, which corresponds to the cut point. In the following Lemma, we show that on a Hypersphere, the $\sin$ function (that appeared in the conditional density function) would smooth out the indicator function, resulting in a smooth vector field.

Lemma 22. *Let $\mathcal{S}^d$ be the unit hypersphere with round metric, $d \geq 2$. Let $x_1 \in \mathcal{S}^d$ and define, for $t \in (0, 1)$ and $x \in \mathcal{S}^d$. Denote $r(x) = d(x, x_1)$ as the radial distance function. Then for integer m with $0 \leq m \leq d - 2$, the function $J(t, x \mid x_1)$ viewed as a function on $(0, 1) \times \mathcal{S}^d$ is C^m.*

Moreover, $v(t, x)$ as a conditional expectation is C^2 in (t, x) for $t \in [0, 1)$. As a result, the solution for flow matching ODE exists and unique.

Proof. Denote $U := \{(t, x) \in (0, 1) \times \mathcal{S}^d : r(x) < (1 - t)\pi\}$. We first show the smoothness on U. Fix $(t_0, x_0) \in U$ where $x_0 \neq x_1$, and write $r_0 := r(x_0) > 0$. Since $r_0 < (1 - t_0)\pi < \pi$, we have $x_0 \neq -x_1$. Moreover, by continuity of $r(\cdot)$ there exists a neighborhood V of (t_0, x_0) such that for all $(t, x) \in V$, $1_{\{r(x) < (1-t)\pi\}} \equiv 1$, and consequently

$$J(t, x \mid x_1) = \frac{1}{1 - t} \Big(\frac{\sin(r(x)/(1 - t))}{\sin r(x)} \Big)^{d-1}.$$

Since $r = d(x, x_1)$ is smooth when $x \notin \{x_1, -x_1\}$, J is smooth as a composition of smooth functions.

It remains to check the smoothness of $J(t, x \mid x_1)$ at point $x = x_1$ ($r = 0$). Introduce normal coordinates at x_1: for x near x_1 write

$$v := \mathrm{Log}_{x_1}(x) \in T_{x_1}\mathcal{S}^d, \qquad r(x) = d(x, x_1) = \|v\|.$$

It suffices to show that under normal coordinates, J viewed as a function of v, is smooth. Notice that $\|v\|$ is not differentiable at $v = 0$, hence functions such as $v \mapsto \sin \|v\|$ are not C^1 at 0. But notice that in J, the "non-smooth" part on $\|v\|$ cancels. By checking smoothness of $\frac{\sin(\|v\|/(1-t))}{\sin \|v\|}$, we conclude the smoothness of J on U.

We extend the smoothness result by checking the boundary $r = (1 - t)\pi$. To prove C^m, fix (t_0, x_0) with $r(x_0) = (1 - t_0)\pi$. Since $r(x_0) \in (0, \pi)$, we have $x \mapsto r(x)$ is smooth near x_0. Set $s(t, x)$ as follows, describing how far will x reach the boundary: $s(t, x) := (1 - t)\pi - r(x)$. On the side $s > 0$ we have $u := r/(1 - t) \in (0, \pi)$ and

$$\sin\Big(\frac{r}{1 - t}\Big) = \sin\Big(\pi - \frac{s}{1 - t}\Big) = \sin\Big(\frac{s}{1 - t}\Big).$$

Therefore, for $s > 0$,

$$J(t,x \mid x_1) = \frac{1}{1-t}\Big(\frac{\sin(s(t,x)/(1-t))}{\sin r(x)}\Big)^{d-1} = s(t,x)^{d-1}(1-t)^{-d}\left(\frac{\frac{\sin(s(t,x)/(1-t))}{s(t,x)/(1-t)}}{\sin r(x)}\right)^{d-1}.$$

In a neighborhood of (t_0, x_0), we have

$$J(t,x \mid x_1) = \big(s(t,x)\big)_+^{d-1}(1-t)^{-d}\left(\frac{\frac{\sin(s(t,x)/(1-t))}{s(t,x)/(1-t)}}{\sin r(x)}\right)^{d-1}, \qquad (s)_+ := \max\{s, 0\}.$$

Since $s \mapsto (s)_+^{d-1}$ is C^m as a one-dimensional function for every $m \le d-2$, and since $s(t,x)$ is smooth in (t,x), the composition $\big(s(t,x)\big)_+^{d-1}$ is C^m in (t,x) for all $m \le d-2$. It follows that J is C^m, as a product of a smooth function and $\big(s(t,x)\big)_+^{d-1}$. □

Remark. *The fact that v is C^2 in (t,x) for $t \in [0,1)$ follows from smoothness of the integrands in rescaled polar coordinates $r = (1-t)u$, and the following properties of Log: (1) Log being uniformly bounded, and (2) singularity of derivatives of $\mathrm{Log}_x(x_1)$ as $x_1 \to -x$ is well controlled under polar coordinates.*

The existence and uniqueness of flow matching ODE follows from classical theory on time-dependent flow, see for example [30, Section 4.1] and [20, Theorem 9.48].

D.1 Auxiliary Lemmas

The following result studies the derivative of a function that only depends on the radial distance. In other words, if a function $\phi(x)$ on M only depends on x through $r(x) = d(x, x_1)$ for some fixed x_1, then we can control its derivative through derivative of r.

Lemma 23. *Fix $x_1 \in \mathcal{S}^d$ and define $r(x) := d(x, x_1)$. Assume $x \notin \mathrm{Cut}\,(x_1)$, so that r is smooth in a neighborhood of x and $\|\mathit{Grad}_x\, r(x)\| = 1$. Let $F\colon (0,\pi) \to \mathbb{R}$ be C^2 and set $\phi(x) := F(r(x))$. Then*

$$\mathit{Grad}_x\, \phi(x) = F'(r(x))\ \mathit{Grad}_x\, r(x), \qquad \|\mathit{Grad}_x\, \phi(x)\| = |F'(r(x))|.$$

Moreover,

$$\|\nabla\, \mathit{Grad}\ \phi(x)\|_{\mathrm{op}} \le |F'(r(x))|\, \|\nabla\, \mathit{Grad}\ r(x)\|_{\mathrm{op}} + |F''(r(x))|.$$

Proof. We apply the chain rule for the Riemannian gradient. For any $u \in T_x S^d$, the differential satisfies

$$d\phi(x)[u] = d(F \circ r)(x)[u] = F'(r(x))\, dr(x)[u].$$

By the definition of the Riemannian gradient,

$$\langle \mathrm{Grad}_x\, \phi, u\rangle = F'(r(x))\, \langle \mathrm{Grad}_x\, r, u\rangle, \qquad \forall u \in T_x S^d,$$

hence

$$\mathrm{Grad}_x\, \phi(x) = F'(r(x))\ \mathrm{Grad}_x\, r(x).$$

Taking norms gives

$$\|\mathrm{Grad}_x\, \phi(x)\| = |F'(r(x))|\, \|\mathrm{Grad}_x\, r(x)\|.$$

Since $x \notin \mathrm{Cut}\,(x_1)$, the distance function satisfies $\|\mathrm{Grad}_x\, r(x)\| = 1$, and therefore

$$\|\mathrm{Grad}_x\, \phi(x)\| = |F'(r(x))|.$$

Next, for any $u \in T_x S^d$,

$$\begin{aligned}\nabla_u\, \mathrm{Grad}\ \phi(x) &= \nabla_u\big(F'(r(x))\ \mathrm{Grad}\ r(x)\big)\\ &= F'(r(x))\, \nabla_u\, \mathrm{Grad}\ r(x) + u\,(F'(r(x)))\, \mathrm{Grad}\ r(x)\end{aligned}$$

$$= F'(r(x))\,\nabla_u \text{ Grad } r(x) + F''(r(x))\,u(r(x))\text{ Grad } r(x)$$
$$= F'(r(x))\,\nabla_u \text{ Grad } r(x) + F''(r(x))\,\langle \text{Grad } r(x), u\rangle \text{ Grad } r(x).$$

Hence, for $\|u\| = 1$,

$$\begin{aligned}\|\nabla_u \text{ Grad } \phi(x)\| &\le |F'(r(x))|\,\|\nabla_u \text{ Grad } r(x)\| + |F''(r(x))|\,|\langle \text{Grad } r(x), u\rangle|\,\|\text{ Grad } r(x)\| \\ &\le |F'(r(x))|\,\|\nabla_u \text{ Grad } r(x)\| + |F''(r(x))|,\end{aligned}$$

using $\|\text{ Grad } r(x)\| = 1$ and $|\langle \text{Grad } r(x), u\rangle| \le \|u\| = 1$. Taking the supremum over $\|u\| = 1$ yields

$$\|\nabla \text{ Grad } \phi(x)\|_{\text{op}} \le |F'(r(x))|\,\|\nabla \text{ Grad } r(x)\|_{\text{op}} + |F''(r(x))|.$$

□

The following result controls the conditional expectation of some certain "functions with singularity", which will be used to establish regularity.

Lemma 24 (Expectation of $\frac{1}{\sin^a u}$.)**.** *Consider $d \ge 3$ and $a = 1, 2$. Let $r = d(x, x_1)$ and set $u = r/(1-t) \in (0, \pi)$. We have*

$$\mathbb{E}[\frac{1}{\sin^a u} \mid X_t = x] \le 2\frac{M_1}{m_1} \qquad \textit{and} \qquad \mathbb{E}[\frac{1}{\sin^a u}] \le 2,$$

where $0 < m_1 \le p_1 \le M_1$.

Proof. We have

$$\begin{aligned}\mathbb{E}\big[\frac{1}{\sin^a u} \mid X_t = x\big] &= \int_{S^d} \frac{1}{\sin^a u}\, p_t(x_1 \mid x)\, dV_g(x_1) \\ &\le \frac{M_1}{m_1 \,\text{Vol}\,(S^d)} \int_{S^d} \frac{1}{\sin^a u}\, J_t(x \mid x_1)\, dV_g(x_1).\end{aligned}$$

Use $r = (1-t)u$ so that $\frac{1}{1-t} dr = du$. Under the polar coordinates around x,

$$\begin{aligned}&\int_{S^d} \frac{1}{\sin^a u}\, J_t(x \mid x_1)\, dV_g(x_1) \\ &\quad= \int_{S^{d-1}} \int_0^\pi \frac{1}{\sin^a u}\, J_t(x \mid x_1)\, (\sin r)^{d-1} dr d\omega \\ &\quad= \text{Vol}\,(S^{d-1}) \int_0^\pi \frac{1}{\sin^a u} \sin(u)^{d-1}\, du \\ &\quad= \text{Vol}\,(S^{d-1}) \int_0^\pi \sin(u)^{d-1-a}\, du.\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{E}\big[\frac{1}{\sin^a u} \mid X_t = x\big] &= \int_{S^d} \frac{1}{\sin^a u}\, p_t(x_1 \mid x)\, dV_g(x_1) \\ &\le \frac{M_1 \,\text{Vol}\,(S^{d-1})}{m_1 \,\text{Vol}\,(S^d)} \int_0^\pi \sin(u)^{d-1-a}\, du. \\ &= \frac{M_1 \,\text{Vol}\,(S^{d-1})}{m_1 \,\text{Vol}\,(S^d)\,\text{Vol}\,(S^{d-1-a})} \,\text{Vol}\,(S^{d-1-a}) \int_0^\pi \sin(u)^{d-1-a}\, du \\ &= \frac{M_1 \,\text{Vol}\,(S^{d-1})}{m_1 \,\text{Vol}\,(S^d)\,\text{Vol}\,(S^{d-1-a})} \,\text{Vol}\,(S^{d-a}).\end{aligned}$$

On the other hand, for the total expectation,

$$\mathbb{E}[\frac{1}{\sin^a u}] = \mathbb{E}[\frac{1}{\sin^a d(X_0, X_1)}] = \mathbb{E}[\mathbb{E}[\frac{1}{\sin^a d(X_0, X_1)} \mid X_1 = x_1]].$$

Since we considered independent coupling between X_0, X_1, we have $\mathbb{E}[\frac{1}{\sin^a d(X_0,X_1)} \mid X_1 = x_1] = \mathbb{E}[\frac{1}{\sin^a d(X_0,x_1)}]$ for any x_1. Since X_0 is the uniform distribution, the expectation has the same value for all x_1. Fixing x_1, using polar coordinates around x_1, we have

$$\begin{aligned}
&\int_{S^d} \frac{1}{\sin^a d(x_0,x_1)} \frac{1}{\operatorname{Vol}(S^d)} \, dV_g(x_0) \\
&\quad = \frac{1}{\operatorname{Vol}(S^d)} \int_{S^{d-1}} \int_0^\pi \frac{1}{\sin^a d(x_0,x_1)} \; (\sin d(x_0,x_1))^{d-1} d(d(x_0,x_1)) d\omega \\
&\quad = \frac{\operatorname{Vol}(S^{d-1})}{\operatorname{Vol}(S^d)} \int_0^\pi \sin(d(x_0,x_1))^{d-1-a} \, d(d(x_0,x_1)) = \frac{\operatorname{Vol}(S^{d-1}) \operatorname{Vol}(S^{d-a})}{\operatorname{Vol}(S^d) \operatorname{Vol}(S^{d-1-a})}.
\end{aligned}$$

Recall

$$\operatorname{Vol}(S^n) = \frac{2\pi^{(n+1)/2}}{\Gamma((n+1)/2)}.$$

Hence

$$\begin{aligned}
\frac{\operatorname{Vol}(S^{d-1})}{\operatorname{Vol}(S^d)} &= \frac{\frac{2\pi^{d/2}}{\Gamma(d/2)}}{\frac{2\pi^{(d+1)/2}}{\Gamma((d+1)/2)}} = \pi^{-1/2} \frac{\Gamma((d+1)/2)}{\Gamma(d/2)}, \\
\frac{\operatorname{Vol}(S^{d-a})}{\operatorname{Vol}(S^{d-1-a})} &= \frac{\frac{2\pi^{(d-a+1)/2}}{\Gamma((d-a+1)/2)}}{\frac{2\pi^{(d-a)/2}}{\Gamma((d-a)/2)}} = \pi^{1/2} \frac{\Gamma((d-a)/2)}{\Gamma((d-a+1)/2)}.
\end{aligned}$$

By Kershaw's inequality, for $d \geq 3$,

$$\frac{\Gamma((d+1)/2)}{\Gamma(d/2)} \leq (d/2 - \frac{1}{2} + \sqrt{\frac{3}{4}})^{\frac{1}{2}},$$

and

$$\frac{\Gamma(d/2)}{\Gamma((d-1)/2)} \geq (\frac{d}{2} - 1 + \frac{1}{2})^{\frac{1}{2}},$$

$$\frac{\Gamma((d-a+1)/2)}{\Gamma((d-a)/2)} \geq (\frac{d-a-1}{2} + \frac{1}{2})^{\frac{1}{2}}.$$

Together,

$$\begin{aligned}
\frac{\operatorname{Vol}(S^{d-1}) \operatorname{Vol}(S^{d-a})}{\operatorname{Vol}(S^d) \operatorname{Vol}(S^{d-1-a})} &= \frac{\Gamma((d+1)/2)\Gamma((d-a)/2)}{\Gamma(d/2)\Gamma((d-a+1)/2)} \leq (\frac{d/2 - \frac{1}{2} + \sqrt{\frac{3}{4}}}{\frac{d-a-1}{2} + \frac{1}{2}})^{\frac{1}{2}} \\
&= (\frac{d-a+a-1+2\sqrt{\frac{3}{4}}}{d-a})^{\frac{1}{2}} \leq (1 + \frac{a-1+2\sqrt{\frac{3}{4}}}{d-a})^{\frac{1}{2}}.
\end{aligned}$$

For $a = 1, 2$, we have $(1 + \frac{a-1+2\sqrt{\frac{3}{4}}}{d-a})^{\frac{1}{2}} \leq 2$. □

The following results (Lemmas 25, 26, 27, and 28) provide bounds on the building blocks that appear in the derivative formulas for $v(t,x)$. For all the results below, we assume $0 < m_1 \leq p_1(x) \leq M_1$.

Lemma 25 (Moment bounds for Grad $_x \log p_t(X_1 \mid x)$.)**.** *We have*

$$\| \mathit{Grad}_x \log J_t(x \mid x_1) \| \leq \frac{2(d-1)}{(1-t)\sin u}.$$

Furthermore,

$$\mathbb{E}[\| \mathit{Grad}_x \log p_t(X_1 \mid x) \| \mid X_t = x] \leq \frac{8(d-1)}{(1-t)} \frac{M_1}{m_1},$$

$$\mathbb{E}[\|\textit{Grad}_x \log J_t(x \mid X_1)\|^2 \mid X_t = x] \le \frac{8(d-1)^2}{(1-t)^2}\frac{M_1}{m_1},$$
$$\mathbb{E}[\|\textit{Grad}_x \log p_t(X_1 \mid x)\|^2 \mid X_t = x] \le \frac{32(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}.$$

Also, using $\mathbb{E}[\frac{1}{\sin^a u}] \le 2$ *from Lemma 24,*

$$\mathbb{E}[\|\textit{Grad}_x \log p_t(X_1 \mid x)\|] \le \frac{8(d-1)}{(1-t)},$$
$$\mathbb{E}[\|\textit{Grad}_x \log J_t(x \mid X_1)\|^2] \le \frac{8(d-1)^2}{(1-t)^2},$$
$$\mathbb{E}[\|\textit{Grad}_x \log p_t(X_1 \mid x)\|^2] \le \frac{32(d-1)^2}{(1-t)^2}.$$

Proof. We first bound $\|\operatorname{Grad}_x \log J_t(x \mid x_1)\|$. Let $r = d(x, x_1)$ and set $u = r/(1-t) \in (0, \pi)$. On the set $r < (1-t)\pi$,

$$\frac{\partial}{\partial r}\log J_t(r) = (d-1)\Big(\frac{1}{1-t}\cot\frac{r}{1-t} - \cot r\Big) = (d-1)\Big(\frac{1}{1-t}\cot u - \cot((1-t)u)\Big).$$

By the Lemma 23,

$$\|\operatorname{Grad}_x \log J_t(x \mid x_1)\| = \Big|\frac{\partial}{\partial r}\log J_t(r)\Big| = (d-1)\Big|\frac{1}{1-t}\cot u - \cot((1-t)u)\Big|.$$

For $u \in (0, \pi)$, we have $|\cot u| \le 1/\sin u$. Since $\sin((1-t)u) \ge (1-t)\sin u$, we have

$$|\cot((1-t)u)| \le \frac{1}{\sin((1-t)u)} \le \frac{1}{(1-t)\sin u}.$$

Therefore for all $u \in (0, \pi)$, we can bound

$$\Big|\frac{1}{1-t}\cot u - \cot((1-t)u)\Big| \le \frac{1}{1-t}|\cot u| + |\cot((1-t)u)| \le \frac{1}{(1-t)\sin u} + \frac{1}{(1-t)\sin u}$$
$$= \frac{2}{(1-t)\sin u}.$$

So

$$\|\operatorname{Grad}_x \log J_t(x \mid x_1)\| \le \frac{2(d-1)}{(1-t)\sin u}.$$

Now we bound $\mathbb{E}\|\operatorname{Grad}_x \log p_t\|$ by $\mathbb{E}\|\operatorname{Grad}_x \log J_t\|$. Let

$$Z_t(x) := \int_{S^d} p_1(z) J_t(x \mid z)\, dV_g(z),$$

so

$$\log p_t(x_1 \mid x) = \log p_1(x_1) + \log J_t(x \mid x_1) - \log Z_t(x).$$

Since p_1 does not depend on x,

$$\operatorname{Grad}_x \log p_t(x_1 \mid x) = \operatorname{Grad}_x \log J_t(x \mid x_1) - \operatorname{Grad}_x \log Z_t(x).$$

Also,

$$\operatorname{Grad}_x Z_t(x) = \int_{S^d} p_1(z) \operatorname{Grad}_x J_t(x \mid z)\, dV_g(z) = \int_{S^d} p_1(z) J_t(x \mid z) \operatorname{Grad}_x \log J_t(x \mid z)\, dV_g(z),$$

hence

$$\operatorname{Grad}_x \log Z_t(x) = \int_{S^d} \operatorname{Grad}_x \log J_t(x \mid z)\, p_t(z \mid x)\, dV_g(z) = \mathbb{E}\big[\operatorname{Grad}_x \log J_t(x \mid X_1) \mid X_t = x\big].$$

Therefore

$$\operatorname{Grad}_x \log p_t(x_1 \mid x) = \operatorname{Grad}_x \log J_t(x \mid x_1) - \mathbb{E}\big[\operatorname{Grad}_x \log J_t(x \mid X_1) \mid X_t = x\big],$$

and by triangle inequality,

$$\mathbb{E}\big[\|\operatorname{Grad}_x \log p_t(X_1 \mid x)\| \mid X_t = x\big] \le 2\,\mathbb{E}\big[\|\operatorname{Grad}_x \log J_t(x \mid X_1)\| \mid X_t = x\big] \le \frac{8(d-1)}{(1-t)}\frac{M_1}{m_1},$$

where we used Lemma 24.

For second moment, we have

$$\mathbb{E}[\|\operatorname{Grad}_x \log p_t(X_1 \mid x)\|^2 \mid X_t = x] \le 4\,\mathbb{E}[\|\operatorname{Grad}_x \log J_t(x \mid X_1)\|^2 \mid X_t = x].$$

Recall $\|\operatorname{Grad}_x \log J_t(x \mid x_1)\| \le \frac{2(d-1)}{(1-t)\sin u}$, using Lemma 24,

$$\mathbb{E}[\|\operatorname{Grad}_x \log J_t(x \mid X_1)\|^2 \mid X_t = x] \le \frac{4(d-1)^2}{(1-t)^2}\mathbb{E}\Big[\frac{1}{\sin^2 u} \mid X_t = x\Big] \le \frac{8(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}.$$

Hence

$$\mathbb{E}[\|\operatorname{Grad}_x \log p_t(X_1 \mid x)\|^2 \mid X_t = x] \le \frac{32(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}.$$

□

Lemma 26 (Moment bounds for $\partial_t \log p_t(X_1 \mid x)$.)**.** *We have*

$$\mathbb{E}\Big[\Big|\frac{\partial}{\partial t}\log p_t(X_1 \mid x)\Big| \mid X_t = x\Big] \le \frac{2}{1-t} + 4\pi\frac{d-1}{1-t}\frac{M_1}{m_1},$$
$$\mathbb{E}\Big[\Big|\frac{\partial}{\partial t}\log p_t(X_1 \mid x)\Big|^2 \mid X_t = x\Big] \le \frac{8}{(1-t)^2} + \frac{16\pi^2(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}.$$

Using $\mathbb{E}[\frac{1}{\sin^a u}] \le 2$ *from Lemma 24,*

$$\mathbb{E}\Big[\Big|\frac{\partial}{\partial t}\log p_t(X_1 \mid x)\Big|\Big] \le \frac{2}{1-t} + 4\pi\frac{d-1}{1-t},$$
$$\mathbb{E}\Big[\Big|\frac{\partial}{\partial t}\log p_t(X_1 \mid x)\Big|^2\Big] \le \frac{8}{(1-t)^2} + \frac{16\pi^2(d-1)^2}{(1-t)^2}.$$

Proof. Let

$$Z_t(x) = \int_{S^d} p_1(z) J_t(x \mid z)\, dV_g(z),$$

so we have $\log p_t(x_1 \mid x) = \log p_1(x_1) + \log J_t(x \mid x_1) - \log Z_t(x)$. Since p_1 does not depend on t,

$$\frac{\partial}{\partial t}\log p_t(x_1 \mid x) = \frac{\partial}{\partial t}\log J_t(x \mid x_1) - \frac{\partial}{\partial t}\log Z_t(x).$$

Moreover, for $d \ge 2$ the function $\partial_t J_t(x \mid z)$ is integrable against $p_1(z)dV_g(z)$, and the map $t \mapsto Z_t(x)$ is C^1 with

$$\frac{\partial}{\partial t}Z_t(x) = \int_{S^d} p_1(z)\frac{\partial}{\partial t}J_t(x \mid z)\, dV_g(z) = \int_{S^d} p_1(z) J_t(x \mid z)\frac{\partial}{\partial t}\log J_t(x \mid z)\, dV_g(z).$$

Dividing by $Z_t(x)$ yields

$$\frac{\partial}{\partial t}\log Z_t(x) = \int_{S^d}\frac{\partial}{\partial t}\log J_t(x \mid z)\, p_t(z \mid x)\, dV_g(z) = \mathbb{E}\Big[\frac{\partial}{\partial t}\log J_t(x \mid X_1) \mid X_t = x\Big].$$

Therefore

$$\frac{\partial}{\partial t}\log p_t(x_1 \mid x) = \frac{\partial}{\partial t}\log J_t(x \mid x_1) - \mathbb{E}\Big[\frac{\partial}{\partial t}\log J_t(x \mid X_1) \mid X_t = x\Big],$$

and by triangle inequality,

$$\mathbb{E}\Big[\Big|\frac{\partial}{\partial t}\log p_t(X_1 \mid x)\Big| \mid X_t = x\Big] \le 2\,\mathbb{E}\Big[\Big|\frac{\partial}{\partial t}\log J_t(x \mid X_1)\Big| \mid X_t = x\Big],$$
$$\mathbb{E}\Big[\Big|\frac{\partial}{\partial t}\log p_t(X_1 \mid x)\Big|^2 \mid X_t = x\Big] \le 4\,\mathbb{E}\Big[\Big|\frac{\partial}{\partial t}\log J_t(x \mid X_1)\Big|^2 \mid X_t = x\Big].$$

We compute $\partial_t \log J_t$ explicitly and bound its conditional expectation. For $r = d(x, x_1) < (1-t)\pi$,

$$\log J_t(x \mid x_1) = -\log(1-t) + (d-1)\Big(\log\sin(r/(1-t)) - \log\sin r\Big).$$

Holding r fixed and differentiating in t gives

$$\frac{\partial}{\partial t}\log J_t(x \mid x_1) = \frac{1}{1-t} + (d-1)\frac{r}{(1-t)^2}\cot(r/(1-t)).$$

Write $u = r/(1-t) \in (0, \pi)$. Then $r = (1-t)u$ and

$$\Big|\frac{\partial}{\partial t}\log J_t(x \mid x_1)\Big| \le \frac{1}{1-t} + \frac{d-1}{1-t}\,|u\cot u|,$$
$$\Big|\frac{\partial}{\partial t}\log J_t(x \mid x_1)\Big|^2 \le 2\frac{1}{(1-t)^2} + 2\frac{(d-1)^2}{(1-t)^2}\,|u\cot u|^2.$$

Hence

$$\mathbb{E}\Big[\Big|\frac{\partial}{\partial t}\log J_t(x \mid X_1)\Big| \mid X_t = x\Big] \le \frac{1}{1-t} + \pi\frac{d-1}{1-t}\,\frac{2M_1}{m_1},$$
$$\mathbb{E}\Big[\Big|\frac{\partial}{\partial t}\log J_t(x \mid X_1)\Big|^2 \mid X_t = x\Big] \le \frac{2}{(1-t)^2} + \frac{2\pi^2(d-1)^2}{(1-t)^2}\,\frac{2M_1}{m_1}.$$

where we used $|u\cot u| \le \frac{\pi}{\sin u}$. Consequently,

$$\mathbb{E}\Big[\Big|\frac{\partial}{\partial t}\log p_t(X_1 \mid x)\Big| \mid X_t = x\Big] \le \frac{2}{1-t} + 4\pi\frac{d-1}{1-t}\,\frac{M_1}{m_1},$$
$$\mathbb{E}\Big[\Big|\frac{\partial}{\partial t}\log p_t(X_1 \mid x)\Big|^2 \mid X_t = x\Big] \le \frac{8}{(1-t)^2} + \frac{16\pi^2(d-1)^2}{(1-t)^2}\,\frac{M_1}{m_1}.$$

□

Lemma 27. *We have*

$$\mathbb{E}\big[\|\nabla_x^2 \log p_t(X_1 \mid x)\|_{\mathrm{op}} \mid X_t = x\big] \le \frac{64(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}.$$

Using $\mathbb{E}[\frac{1}{\sin^a u}] \le 2$ *from Lemma 24,*

$$\mathbb{E}\big[\|\nabla_x^2 \log p_t(X_1 \mid x)\|_{\mathrm{op}}\big] \le \frac{64(d-1)^2}{(1-t)^2}.$$

Proof. Recall that

$$\mathsf{Grad}_x \log p_t(x_1 \mid x) = \mathsf{Grad}_x \log J_t(x \mid x_1) - \mathsf{Grad}_x \log Z_t(x),$$

Take derivative again, we obtain

$$\nabla\,\mathsf{Grad}_x \log p_t(x_1 \mid x) = \nabla\,\mathsf{Grad}_x \log J_t(x \mid x_1) - \nabla\,\mathsf{Grad}_x \log Z_t(x),$$

Recall

$$\mathsf{Grad}_x \log Z_t(x) = \mathbb{E}\big[\mathsf{Grad}_x \log J_t(x \mid X_1) \mid X_t = x\big].$$

Hence

$$\nabla_u\,\mathsf{Grad}_x \log Z_t(x) = \nabla_u \mathbb{E}\big[\mathsf{Grad}_x \log J_t(x \mid X_1) \mid X_t = x\big]$$

$$=\mathbb{E}\big[\nabla_u \operatorname{Grad}_x \log J_t(x \mid X_1) \mid X_t = x\big] + \mathbb{E}\big[\operatorname{Grad}_x \log J_t(x \mid X_1) \nabla_u \log p_t(x_1 \mid x) \mid X_t = x\big],$$

and we can bound

$$\begin{aligned}
&\|\nabla \operatorname{Grad}_x \log Z_t(x)\|_{\mathrm{op}} \\
&\le \mathbb{E}\big[\|\nabla^2 \log J_t(x \mid X_1)\|_{\mathrm{op}} \mid X_t = x\big] \\
&\quad + \sqrt{\mathbb{E}\big[\|\operatorname{Grad}_x \log J_t(x \mid X_1)\|^2 \mid X_t = x\big] \mathbb{E}\big[\|\nabla \log p_t(x_1 \mid x)\|^2 \mid X_t = x\big]} \\
&\le \mathbb{E}\big[\|\nabla^2 \log J_t(x \mid X_1)\|_{\mathrm{op}} \mid X_t = x\big] + \frac{16(d-1)^2 M_1}{(1-t)^2 m_1}.
\end{aligned}$$

To control $\|\nabla^2 \log J_t(x \mid X_1)\|_{\mathrm{op}}$, we consider

$$\begin{aligned}
\frac{\partial^2}{\partial r^2} \log J_t(x \mid X_1) &= \frac{\partial}{\partial r}(d-1)\Big(\frac{1}{1-t}\cot\frac{r}{1-t} - \cot r\Big) = (d-1)\Big(-\frac{1}{(1-t)^2}\frac{1}{\sin^2 \frac{r}{1-t}} + \frac{1}{\sin^2 r}\Big) \\
&= (d-1)\Big(\frac{1}{\sin^2(1-t)u} - \frac{1}{(1-t)^2}\frac{1}{\sin^2 u}\Big).
\end{aligned}$$

Applying Lemma 23 with $\phi = \log J_t(r)$, we have

$$\begin{aligned}
\|\nabla \operatorname{Grad} \log J_t(r)\|_{\mathrm{op}} &\le |\log J_t'(r(x))| \|\nabla \operatorname{Grad} r(x)\|_{\mathrm{op}} + |\log J_t''(r(x))| \\
&\le |\log J_t'(r(x))| \cot r + |\log J_t''(r(x))|,
\end{aligned}$$

where $\|\nabla \operatorname{Grad} r(x)\|_{\mathrm{op}} \le |\cot r|$ by **(author?)** [21, Proposition 11.3]

Now using

$$\begin{aligned}
\frac{\partial}{\partial r} \log J_t(x \mid X_1) &= (d-1)\Big(\frac{1}{1-t}\cot\frac{r}{1-t} - \cot r\Big), \\
\frac{\partial^2}{\partial r^2} \log J_t(x \mid X_1) &= (d-1)\Big(\frac{1}{\sin^2(1-t)u} - \frac{1}{(1-t)^2}\frac{1}{\sin^2 u}\Big),
\end{aligned}$$

we obtain

$$\|\nabla^2 \log J_t(x \mid X_1)\|_{\mathrm{op}} \le \frac{4(d-1)}{(1-t)^2 \sin^2 u},$$

hence

$$\mathbb{E}[\|\nabla^2 \log J_t(x \mid X_1)\| \mid X_t = x] \le \frac{16(d-1)}{(1-t)^2}\frac{M_1}{m_1}.$$

Now we bound $\mathbb{E}[\|\nabla_x^2 \log p_t(X_1 \mid x)\|_{\mathrm{op}} \mid X_t = x]$. Recall

$$\begin{aligned}
\nabla \operatorname{Grad}_x \log p_t(x_1 \mid x) &= \nabla \operatorname{Grad}_x \log J_t(x \mid x_1) - \nabla \operatorname{Grad}_x \log Z_t(x), \\
\|\nabla \operatorname{Grad}_x \log Z_t(x)\|_{\mathrm{op}} &\le \mathbb{E}\big[\|\nabla^2 \log J_t(x \mid X_1)\|_{\mathrm{op}} \mid X_t = x\big] + \frac{16(d-1)^2 M_1}{(1-t)^2 m_1}.
\end{aligned}$$

By triangle inequality, we obtain

$$\mathbb{E}\big[\|\nabla_x^2 \log p_t(X_1 \mid x)\|_{\mathrm{op}} \mid X_t = x\big] \le 2\,\mathbb{E}\big[\|\nabla_x^2 \log J_t(x \mid X_1)\|_{\mathrm{op}} \mid X_t = x\big] + \frac{16(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}.$$

Finally, using the bound

$$\mathbb{E}\big[\|\nabla_x^2 \log J_t(x \mid X_1)\|_{\mathrm{op}} \mid X_t = x\big] \le \frac{16(d-1)}{(1-t)^2}\frac{M_1}{m_1},$$

we conclude

$$\mathbb{E}\big[\|\nabla_x^2 \log p_t(X_1 \mid x)\|_{\mathrm{op}} \mid X_t = x\big] \le \frac{32(d-1)}{(1-t)^2}\frac{M_1}{m_1} + \frac{16(d-1)^2}{(1-t)^2}\frac{M_1}{m_1} < \frac{64(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}.$$

□

Lemma 28. *We have*

$$\mathbb{E}[\|\partial_t \textit{Grad}_x \log p_t(X_1 \mid x)\| \mid X_t = x] \le \frac{16\pi^2(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}.$$

Using $\mathbb{E}[\frac{1}{\sin^a u}] \le 2$ *from Lemma 24,*

$$\mathbb{E}[\|\partial_t \textit{Grad}_x \log p_t(X_1 \mid x)\|] \le \frac{16\pi^2(d-1)^2}{(1-t)^2}.$$

Proof. We have

$$\begin{aligned}\operatorname{Grad}_x \log p_t(x_1 \mid x) &= \operatorname{Grad}_x \log J_t(x \mid x_1) - \operatorname{Grad}_x \log Z_t(x)\\ &= \operatorname{Grad}_x \log J_t(x \mid x_1) - \mathbb{E}\big[\operatorname{Grad}_x \log J_t(x \mid X_1) \mid X_t = x\big],\end{aligned}$$

so we have

$$\begin{aligned}\|\partial_t \operatorname{Grad}_x \log p_t(x_1 \mid x)\| \le{}& \|\partial_t \operatorname{Grad}_x \log J_t(x \mid x_1)\| + \mathbb{E}\big[\|\partial_t \operatorname{Grad}_x \log J_t(x \mid X_1)\| \mid X_t = x\big]\\ &+ \mathbb{E}[\|\operatorname{Grad}_x \log J_t(x \mid X_1)\|\,|\partial_t \log p_t(X_1 \mid x)| \mid X_t = x].\end{aligned}$$

A direct differentiation at fixed r gives

$$\begin{aligned}\|\partial_t \operatorname{Grad}_x \log J_t(x \mid x_1)\| &= \|\partial_t \frac{\partial}{\partial r} \log J_t(r)\operatorname{Grad}\, r(x)\| = (d-1)\Big\|\partial_t \frac{1}{1-t}\cot u - \cot((1-t)u)\Big\|\\ &= (d-1)\Big\|\frac{1}{(1-t)^2}\cot u - \frac{u}{\sin^2(1-t)u}\Big\|\\ &\le \frac{d-1}{(1-t)^2}\frac{2\pi}{\sin^2 u}.\end{aligned}$$

Taking expectation and using $\mathbb{E}[1/\sin^a u] \le 2M_1/m_1$ for $a = 1, 2$,

$$\mathbb{E}[\|\partial_t \operatorname{Grad}_x \log J_t(x \mid X_1)\| \mid X_t = x] \le \frac{4\pi(d-1)}{(1-t)^2}\frac{M_1}{m_1},$$

Recall that

$$\begin{aligned}\mathbb{E}\Big[\Big|\partial_t \log p_t(X_1 \mid x)\Big|^2 \mid X_t = x\Big] &\le 4\mathbb{E}\Big[\Big|\partial_t \log J_t\Big|^2 \mid X_t = x\Big]\\ &\le 4\mathbb{E}\Big[\frac{(d-1)^2\pi^2}{(1-t)^2}\frac{1}{\sin^2 u} \mid X_t = x\Big] \le \frac{8\pi^2(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}.\end{aligned}$$

Together with $\|\operatorname{Grad}_x \log J_t(x \mid x_1)\| \le \frac{2(d-1)}{(1-t)\sin u}$, by Cauchy-Schwarz,

$$\begin{aligned}&\mathbb{E}[\|\operatorname{Grad}_x \log J_t(x \mid X_1)\|\,|\partial_t \log p_t(X_1 \mid x)| \mid X_t = x]\\ &\le \Big(\mathbb{E}[\|\operatorname{Grad}_x \log J_t(x \mid X_1)\|^2 \mid X_t = x]\Big)^{1/2}\Big(\mathbb{E}[|\partial_t \log p_t(X_1 \mid x)|^2 \mid X_t = x]\Big)^{1/2}\\ &\le \sqrt{\frac{8\pi(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}\frac{8\pi^2(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}} \le \frac{8\pi^2(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}.\end{aligned}$$

Thus

$$\mathbb{E}[\|\partial_t \operatorname{Grad}_x \log p_t(X_1 \mid x)\| \mid X_t = x] \le \frac{16\pi^2(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}.$$

□

D.2 Regularity for Flow Matching Vector Field

In this section, we show that both the spatial derivative (Lemma 29) and time derivative (Lemma 30) of $v(t,x)$ are uniformly bounded.

Lemma 29. *Assume $d \geq 3$ and $t \in (0,1)$. Let p_1 be a smooth density on S^d such that*

$$0 < m_1 \leq p_1(z) \leq M_1 < \infty.$$

For all unit tangent vector w, we have

$$\|\nabla_w v(t,x)\| \leq \frac{12\pi M_1(d-1)}{m_1(1-t)}.$$

Hence $v(t,x)$ is L-Lipschitz with $L = \frac{12\pi M_1(d-1)}{m_1(1-t)}$. Replacing the conditional expectation bounds by expectation bounds in the proof (based on 24), we obtain

$$\mathbb{E}[\|\nabla_w v(t,x)\|] \leq \frac{12\pi(d-1)}{(1-t)}.$$

Proof. Fix $t < 1$ and $x \in M$, and let $w \in T_xM$ be a unit tangent vector. We consider the covariant derivative of v in the direction w:

$$\begin{aligned}
\nabla_w v(t,x) &= \frac{1}{1-t}\nabla_w \int_M \operatorname{Log}_x(x_1)\, p_t(x_1|x)\, dV_g(x_1) \\
&= \frac{1}{1-t}\int_M \nabla_w\big(\operatorname{Log}_x(x_1)\big)\, p_t(x_1|x)\, dV_g(x_1) \\
&\quad + \frac{1}{1-t}\int_M \operatorname{Log}_x(x_1)\, \langle \operatorname{Grad}_x p_t(x_1|x), w\rangle\, dV_g(x_1).
\end{aligned}$$

Define

$$\begin{aligned}
P(t,x) &:= \frac{1}{1-t}\int_{S^d} \operatorname{Log}_x(x_1)\langle \operatorname{Grad}_x p_t(x_1 \mid x), w\rangle\, dV_g(x_1), \\
G(t,x) &:= \frac{1}{1-t}\int_{S^d} \|\nabla_w \operatorname{Log}_x(x_1)\| p_t(x_1 \mid x)\, dV_g(x_1).
\end{aligned}$$

We show that $\|P(t,x)\| \leq \frac{8\pi M_1(d-1)}{m_1(1-t)}, \forall \|w\| = 1$. We first write $\|P(t,x)\|$ as a conditional moment of $\|\operatorname{Grad}_x \log p_t\|$. Using $\operatorname{Grad}_x p_t = p_t \operatorname{Grad}_x \log p_t$ and $\|\operatorname{Log}_x(x_1)\| = d(x,x_1)$,

$$\begin{aligned}
\|P(t,x)\| &\leq \frac{1}{1-t}\int_{S^d} \|\operatorname{Log}_x(x_1)\|\, \|\operatorname{Grad}_x p_t(x_1 \mid x)\|\, dV_g(x_1) \\
&= \frac{1}{1-t}\int_{S^d} \|\operatorname{Log}_x(x_1)\|\, p_t(x_1 \mid x)\, \|\operatorname{Grad}_x \log p_t(x_1 \mid x)\|\, dV_g(x_1) \\
&= \frac{1}{1-t}\mathbb{E}\big[\|\operatorname{Log}_x(X_1)\|\, \|\operatorname{Grad}_x \log p_t(X_1 \mid x)\| \mid X_t = x\big].
\end{aligned}$$

On the support of $p_t(\cdot \mid x)$ we have $d(x,X_1) < (1-t)\pi$, hence

$$\|\operatorname{Log}_x(X_1)\| \leq (1-t)\pi,$$

so

$$\|P(t,x)\| \leq \pi\, \mathbb{E}\big[\|\operatorname{Grad}_x \log p_t(X_1 \mid x)\| \mid X_t = x\big].$$

Using Lemma 25,

$$\|P(t,x)\| \leq \frac{8\pi M_1(d-1)}{m_1(1-t)}.$$

Next, we show that $G(t,x) \leq 4\pi\frac{M_1}{m_1(1-t)}$.

Notice that $\nabla_w \operatorname{Log}_x(x_1)$ can be related to a Hessian. Denote $r = d(x, x_1) \in (0, \pi)$, it is easy to show (see, for example, [1, Appendix B, Proof of Lemma 2])

$$\|\nabla_w \operatorname{Log}_x(x_1)\| \le 1 + |r \cot(r)|.$$

Now we bound the cot term. On the support of $p_t(\cdot \mid x)$ we have $r < (1-t)\pi$. Let $u = r/(1-t) \in (0, \pi)$, so $r = (1-t)u$. Using $|\cot r| \le 1/\sin r$ and the concavity bound $\sin((1-t)u) \ge (1-t)\sin u$,

$$|r \cot r| \le \frac{r}{\sin r} = \frac{(1-t)u}{\sin((1-t)u)} \le \frac{(1-t)u}{(1-t)\sin u} = \frac{u}{\sin u} \le \frac{\pi}{\sin u}.$$

Therefore for $r < (1-t)\pi$,

$$\|\nabla_w \operatorname{Log}_x(x_1)\| \le 1 + \frac{\pi}{\sin u} < \frac{2\pi}{\sin u}.$$

Using the same technique as before,

$$\int_{S^d} \|\nabla_w \operatorname{Log}_x(x_1)\| \, p_t(x_1 \mid x) \, dV_g(x_1) \le 4\pi \frac{M_1}{m_1}.$$

We conclude that Multiplying by $(1-t)^{-1}$ yields

$$G(t, x) \le 4\pi \frac{M_1}{m_1(1-t)}.$$

□

Remark. *In the above proof, by taking w over an orthonormal basis $\{e_i\}_{i=1}^d$, we obtain $\mathbb{E}\|\nabla v(t, X_t)\|_{\mathrm{HS}}^2 \le 80\pi^2(d-1)^2/(1-t)^2$.*

Lemma 30. *Fix $d \ge 3$ and $t \in (0, 1)$. Let p_1 be a smooth density on S^d satisfying*

$$0 < m_1 \le p_1(z) \le M_1 < \infty.$$

For

$$v(t, x) = \frac{1}{1-t} \int_{S^d} \mathit{Log}_x(x_1) \, p_t(x_1 \mid x) \, dV_g(x_1),$$

we have for every $x \in S^d$,

$$\Big\| \frac{d}{dt} v(t, x_t) \Big\| \le \frac{20\pi^2 d}{1-t} \frac{M_1}{m_1}.$$

Replacing the conditional expectation bounds by expectation bounds in the proof (based on 24), we obtain

$$\mathbb{E}\Big[\Big\| \frac{d}{dt} v(t, x) \Big\| \Big] \le \frac{20\pi^2 d}{1-t}.$$

Proof. Observe that on the support of $p_t(\cdot \mid x)$ we have $d(x, x_1) < (1-t)\pi$, hence

$$\|\operatorname{Log}_x(x_1)\| = d(x, x_1) \le (1-t)\pi.$$

Therefore

$$\|v(t, x)\| \le \frac{1}{1-t} \int_{S^d} \|\operatorname{Log}_x(x_1)\| \, p_t(x_1 \mid x) \, dV_g(x_1) \le \frac{1}{1-t}(1-t)\pi = \pi.$$

We differentiate $v(t, x)$ in t. Since $\operatorname{Log}_x(x_1)$ does not depend on t,

$$\begin{aligned} \partial_t v(t, x) &= \frac{d}{dt}\Big(\frac{1}{1-t}\Big) \int_{S^d} \operatorname{Log}_x(x_1) \, p_t(x_1 \mid x) \, dV_g(x_1) + \frac{1}{1-t} \int_{S^d} \operatorname{Log}_x(x_1) \, \frac{\partial}{\partial t} p_t(x_1 \mid x) \, dV_g(x_1) \\ &= \frac{1}{(1-t)^2} \int_{S^d} \operatorname{Log}_x(x_1) \, p_t(x_1 \mid x) \, dV_g(x_1) + \frac{1}{1-t} \int_{S^d} \operatorname{Log}_x(x_1) \, \frac{\partial}{\partial t} p_t(x_1 \mid x) \, dV_g(x_1). \end{aligned}$$

Using

$$\int_{S^d} \mathrm{Log}_x(x_1)\, p_t(x_1 \mid x)\, dV_g(x_1) = (1-t)\, v(t,x),$$

we have

$$\partial_t v(t,x) = \frac{1}{1-t} v(t,x) + \frac{1}{1-t} \int_{S^d} \mathrm{Log}_x(x_1)\, \frac{\partial}{\partial t} p_t(x_1 \mid x)\, dV_g(x_1).$$

Since $\int_{S^d} p_t(x_1 \mid x)\, dV_g(x_1) = 1$, we have

$$\int_{S^d} \frac{\partial}{\partial t} p_t(x_1 \mid x)\, dV_g(x_1) = 0.$$

Also, wherever $p_t(x_1 \mid x) > 0$,

$$\frac{\partial}{\partial t} p_t(x_1 \mid x) = p_t(x_1 \mid x)\, \frac{\partial}{\partial t} \log p_t(x_1 \mid x).$$

Hence

$$\begin{aligned}\Big\| \int_{S^d} \mathrm{Log}_x(x_1)\, \frac{\partial}{\partial t} p_t(x_1 \mid x)\, dV_g(x_1) \Big\| &\le \int_{S^d} \| \mathrm{Log}_x(x_1) \|\, p_t(x_1 \mid x) \Big| \frac{\partial}{\partial t} \log p_t(x_1 \mid x) \Big|\, dV_g(x_1) \\ &\le (1-t)\pi\, \mathbb{E}\Big[\Big| \frac{\partial}{\partial t} \log p_t(X_1 \mid x) \Big| \mid X_t = x \Big].\end{aligned}$$

Therefore

$$\Big\| \partial_t v(t,x) \Big\| \le \frac{\pi}{1-t} + \pi\, \mathbb{E}\Big[\Big| \frac{\partial}{\partial t} \log p_t(X_1 \mid x) \Big| \mid X_t = x \Big].$$

Using Lemma 26,

$$\begin{aligned}\Big\| \partial_t v(t,x) \Big\| &\le \frac{\pi}{1-t} + \pi\, \mathbb{E}\Big[\Big| \frac{\partial}{\partial t} \log p_t(X_1 \mid x) \Big| \mid X_t = x \Big] \\ &\le \frac{\pi}{1-t} + 2\pi \Big(\frac{1}{1-t} + \frac{d-1}{1-t} \frac{2\pi M_1}{m_1} \Big) \le \frac{1}{1-t} \Big(3\pi + 4\pi^2 \frac{d M_1}{m_1} \Big).\end{aligned}$$

Since $\|v(t,x)\| \le \pi$, Lemma 29 gives

$$\|\nabla_{v(t,x)} v(t,x)\| \le \|\nabla v(t,x)\|_{\mathrm{op}} \|v(t,x)\| \le \frac{12\pi^2 (d-1)}{1-t} \frac{M_1}{m_1}.$$

Together with the preceding bound for $\partial_t v(t,x)$, we obtain

$$\| \frac{d}{dt} v(t, x_t) \| = \| \partial_t v(t,x) + \nabla_{v(t,x)} v(t,x) \| \le \frac{20\pi^2 d}{1-t} \frac{M_1}{m_1}.$$

□

D.3 Regularity for Divergence

In this section, we show that both the gradient (Lemma 31) and the time derivative (Lemma 32) of Div $v(t,x)$ are uniformly bounded.

We first observe, under our notation $r = d(x, x_1)$ where x stands for x_t, we have

$$r|\cot r| \begin{cases} \le 1, \text{ if } t \ge \frac{1}{2}; \\ < \pi \frac{1}{(1-t)\sin u} \le \frac{2\pi}{\sin u}, \text{ if } t < \frac{1}{2}. \end{cases}$$

Lemma 31. *Assume $d \ge 3$ and $t \in (0,1)$. Let p_1 be a smooth density on S^d such that*

$$0 < m_1 \le p_1(z) \le M_1 < \infty.$$

Then for every $x \in S^d$ we have

$$\| \textit{Grad}_x \textit{ Div } v(t,x)\| \le \frac{128\pi(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}.$$

Replacing the conditional expectation bounds by expectation bounds in the proof (based on 24), we obtain

$$\mathbb{E}[\| \textit{Grad}_x \textit{ Div } v(t,x)\|] \le \frac{128\pi(d-1)^2}{(1-t)^2}.$$

Proof. Write

$$v(t,x) = \frac{1}{1-t}\int_{S^d} \mathrm{Log}_x(x_1)\, p_t(x_1 \mid x)\, dV_g(x_1).$$

Using $\mathrm{Div}\,(fW) = \langle \mathrm{Grad}\ f, W\rangle + f\,\mathrm{Div}\ W$ with $f = p_t(\cdot \mid x)$ and $W = \mathrm{Log}_x(\cdot)$,

$$\begin{aligned}
\mathrm{Div}\ v(t,x) &= \frac{1}{1-t}\int_{S^d} \mathrm{Div}_x \mathrm{Log}_x(x_1)\, p_t(x_1 \mid x)\, dV_g(x_1) \\
&\quad + \frac{1}{1-t}\int_{S^d} \langle \mathrm{Grad}_x\, p_t(x_1 \mid x), \mathrm{Log}_x(x_1)\rangle\, dV_g(x_1) \\
&= \frac{1}{1-t}\mathbb{E}[\mathrm{Div}_x \mathrm{Log}_x(X_1) \mid X_t = x] \\
&\quad + \frac{1}{1-t}\mathbb{E}[\langle \mathrm{Grad}_x \log p_t(X_1 \mid x), \mathrm{Log}_x(X_1)\rangle \mid X_t = x].
\end{aligned}$$

Differentiate this identity in x along a unit vector $\xi \in T_xS^d$ and take norms. We take gradient, and obtain (for a smooth integrand $F(x, x_1)$),

$$\mathrm{Grad}_x\, \mathbb{E}[F(x,X_1) \mid X_t = x] = \mathbb{E}[\mathrm{Grad}_x\, F(x,X_1) \mid X_t = x] + \mathbb{E}[F(x,X_1)\, \mathrm{Grad}_x \log p_t(X_1 \mid x) \mid X_t = x].$$

Applying this formula and using triangle inequality yields

$$\| \mathrm{Grad}_x\, \mathrm{Div}\ v(t,x)\| \le \frac{1}{1-t}\Big(T_1 + T_2 + T_3 + T_4\Big),$$

where

$$\begin{aligned}
T_1 &= \mathbb{E}[\| \mathrm{Grad}_x \mathrm{Div}_x \mathrm{Log}_x(X_1)\| \mid X_t = x], \\
T_2 &= \mathbb{E}[|\mathrm{Div}_x \mathrm{Log}_x(X_1)|\, \| \mathrm{Grad}_x \log p_t(X_1 \mid x)\| \mid X_t = x], \\
T_3 &= \mathbb{E}[\| \mathrm{Grad}_x \langle \mathrm{Grad}_x \log p_t(X_1 \mid x), \mathrm{Log}_x(X_1)\rangle\| \mid X_t = x], \\
T_4 &= \mathbb{E}[|\langle \mathrm{Grad}_x \log p_t(X_1 \mid x), \mathrm{Log}_x(X_1)\rangle|\, \| \mathrm{Grad}_x \log p_t(X_1 \mid x)\| \mid X_t = x].
\end{aligned}$$

Now we compute $\mathrm{Div}_x \mathrm{Log}_x(x_1)$ and its x-gradient. Let $r = d(x, x_1) \in (0, \pi)$. Notice that

$$\mathrm{Log}_x(x_1) = -\,\mathrm{Grad}_x \frac{1}{2}r^2,$$

hence

$$\mathrm{Div}_x \mathrm{Log}_x(x_1) = -\,\mathrm{Div}_x(r\,\mathrm{Grad}\ r) = -r\Delta r - \langle \mathrm{Grad}\ r, \mathrm{Grad}\ r\rangle = -(d-1)r\cot r - 1,$$

where we used $\Delta r = (d-1)\cot r$ (see for example [21, Theorem 11.11]) and $\| \mathrm{Grad}\ r\|^2 = 1$. Differentiate in x using that r is a radial function and $\| \mathrm{Grad}\ r\| = 1$:

$$\mathrm{Grad}_x \mathrm{Div}_x \mathrm{Log}_x(x_1) = -(d-1)\big(\cot r - r\csc^2 r\big)\,\mathrm{Grad}_x\, r,$$

which implies

$$\| \mathrm{Grad}_x \mathrm{Div}_x \mathrm{Log}_x(x_1)\| = (d-1)\big|\cot r - r\csc^2 r\big| = (d-1)\Big|\cot r - r\frac{1}{\sin^2 r}\Big|.$$

We bound $T_1 = \mathbb{E}[\|\operatorname{Grad}_x \operatorname{Div}_x \operatorname{Log}_x(X_1)\| \mid X_t = x]$. Using $r = (1-t)u$ and $\sin((1-t)u) \geq (1-t)\sin u$ (concavity of sin on $[0,\pi]$),

$$|\cot r| \leq \frac{1}{\sin r} \leq \frac{1}{(1-t)\sin u}, \qquad r\frac{1}{\sin^2 r} \leq \frac{(1-t)\pi}{\sin^2 r} \leq \frac{\pi}{(1-t)\sin^2 u}.$$

Hence

$$\|\operatorname{Grad}_x \operatorname{Div}_x \operatorname{Log}_x(x_1)\| \leq (d-1)\Big(\frac{1}{(1-t)\sin u} + \frac{\pi}{(1-t)\sin^2 u}\Big).$$

Therefore

$$T_1 \leq (d-1)\frac{1}{1-t}\frac{M_1}{m_1}(2\pi+2).$$

We bound $T_2 = \mathbb{E}[|\operatorname{Div}_x \operatorname{Log}_x(X_1)| \, \|\operatorname{Grad}_x \log p_t(X_1 \mid x)\| \mid X_t = x]$. Recall $|\operatorname{Div}_x \operatorname{Log}_x(x_1)| \leq 1 + (d-1)r|\cot r|$, so we have

$$|\operatorname{Div}_x \operatorname{Log}_x(x_1)| \leq (d-1)\frac{4\pi}{\sin u}.$$

Thus by Cauchy-Schwarz,

$$T_2 \leq \Big(\mathbb{E}[|\operatorname{Div}_x \operatorname{Log}_x(X_1)|^2 \mid X_t = x]\Big)^{1/2}\Big(\mathbb{E}[\|\operatorname{Grad}_x \log p_t(X_1 \mid x)\|^2 \mid X_t = x]\Big)^{1/2}.$$

Using $(a+b)^2 \leq 2a^2 + 2b^2$,

$$\begin{aligned}\mathbb{E}[|\operatorname{Div}_x \operatorname{Log}_x(X_1)|^2 \mid X_t = x] &\leq 4(d-1)^2\pi^2\mathbb{E}\Big[\frac{1}{\sin^2 u} \mid X_t = x\Big] \\ &\leq 8(d-1)^2\pi^2\frac{M_1}{m_1}.\end{aligned}$$

Also recall

$$\mathbb{E}[\|\operatorname{Grad}_x \log p_t(X_1 \mid x)\|^2 \mid X_t = x] \leq \frac{32(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}.$$

Hence we obtain

$$\begin{aligned}T_2 &\leq (8(d-1)^2\pi^2\frac{M_1}{m_1})^{\frac{1}{2}}(\frac{32(d-1)^2}{(1-t)^2}\frac{M_1}{m_1})^{\frac{1}{2}} \\ &\leq \frac{16\pi M_1}{m_1}\frac{(d-1)^2}{(1-t)}.\end{aligned}$$

We bound $T_3 = \mathbb{E}[\|\operatorname{Grad}_x\langle \operatorname{Grad}_x \log p_t(X_1 \mid x), \operatorname{Log}_x(X_1)\rangle\| \mid X_t = x]$. We use the product rule: denote $g(x) = \langle g_1(x), g_2(x)\rangle$. For g_1, g_2, we have (viewing as directional derivative along u and use compatiblity)

$$\langle \operatorname{Grad}_x\langle g_1(x), g_2(x)\rangle, u\rangle = \nabla_u\langle g_1(x), g_2(x)\rangle = \langle \nabla_u g_1(x), g_2(x)\rangle + \langle g_1(x), \nabla_u g_2(x)\rangle.$$

Hence we can write

$$\begin{aligned}&\|\operatorname{Grad}_x\langle \operatorname{Grad}_x \log p_t(x_1 \mid x), \operatorname{Log}_x(x_1)\rangle\| \\ =&\|\nabla \operatorname{Grad}_x \log p_t(x_1 \mid x)\|_{\mathrm{op}} \|\operatorname{Log}_x(x_1)\| + \|\nabla \operatorname{Log}_x(x_1)\|_{\mathrm{op}} \|\operatorname{Grad}_x \log p_t(x_1 \mid x)\|,\end{aligned}$$

so

$$\begin{aligned}T_3 \leq\ &\mathbb{E}[\|\nabla \operatorname{Log}_x(X_1)\| \, \|\operatorname{Grad}_x \log p_t(X_1 \mid x)\| \mid X_t = x] \\ &+ \mathbb{E}[\|\nabla^2 \log p_t(X_1 \mid x)\| \, \|\operatorname{Log}_x(X_1)\| \mid X_t = x].\end{aligned}$$

On the support of $p_t(\cdot \mid x)$, $\|\operatorname{Log}_x(X_1)\| \leq (1-t)\pi$, hence

$$\mathbb{E}[\|\nabla^2 \log p_t(X_1 \mid x)\| \, \|\operatorname{Log}_x(X_1)\| \mid X_t = x] \leq (1-t)\pi\,\mathbb{E}[\|\nabla^2 \log p_t(X_1 \mid x)\| \mid X_t = x].$$

By Lemma 27,

$$\mathbb{E}[\|\nabla^2 \log p_t(X_1 \mid x)\| \, \|\operatorname{Log}_x(X_1)\| \mid X_t = x] \le \frac{64\pi(d-1)^2}{1-t}\frac{M_1}{m_1}.$$

Also, on S^d, we have that

$$\|\nabla \operatorname{Log}_x(x_1)\| \le 1 + r|\cot r| \begin{cases} \le 2, \text{ if } t \ge \frac{1}{2}; \\ < \pi\frac{2}{(1-t)\sin u} \le \frac{4\pi}{\sin u}, \text{ if } t < \frac{1}{2}. \end{cases}$$

So by Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}[\|\nabla \operatorname{Log}_x(X_1)\| \, \|\operatorname{Grad}_x \log p_t(X_1 \mid x)\| \mid X_t = x] &\le \Big(\frac{32(d-1)^2}{(1-t)^2}\frac{M_1}{m_1} 16\pi^2 \frac{M_1}{m_1}\Big)^{\frac{1}{2}} \\ &\le \frac{32\pi M_1}{m_1}\frac{(d-1)}{(1-t)}.\end{aligned}$$

Hence we get

$$T_3 \le \frac{32\pi M_1}{m_1}\frac{(d-1)}{(1-t)} + \frac{64\pi(d-1)^2}{1-t}\frac{M_1}{m_1} \le \frac{64\pi M_1}{m_1}\frac{(d-1)^2}{(1-t)}.$$

We bound $T_4 = \mathbb{E}[|\langle \operatorname{Grad}_x \log p_t(X_1 \mid x), \operatorname{Log}_x(X_1)\rangle| \, \|\operatorname{Grad}_x \log p_t(X_1 \mid x)\| \mid X_t = x]$. We have

$$\begin{aligned}T_4 &\le \mathbb{E}[\|\operatorname{Log}_x(X_1)\| \, \|\operatorname{Grad}_x \log p_t(X_1 \mid x)\|^2 \mid X_t = x] \\ &\le (1-t)\pi \, \mathbb{E}[\|\operatorname{Grad}_x \log p_t(X_1 \mid x)\|^2 \mid X_t = x] \\ &\le (1-t)\pi \cdot \frac{32(d-1)^2}{(1-t)^2}\frac{M_1}{m_1} = \frac{32\pi(d-1)^2}{1-t}\frac{M_1}{m_1}.\end{aligned}$$

Together,

$$\begin{aligned}&\|\operatorname{Grad}_x \operatorname{Div} v(t,x)\| \\ \le&\frac{1}{1-t}\Big(T_1 + T_2 + T_3 + T_4\Big) \\ \le&\frac{1}{(1-t)}\frac{M_1}{m_1}\Big((d-1)\frac{1}{1-t}(2\pi+2) + 16\pi\frac{(d-1)^2}{1-t} + 64\pi\frac{(d-1)^2}{1-t} + \frac{32\pi(d-1)^2}{1-t}\Big) \\ \le&\frac{128\pi(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}.\end{aligned}$$

□

Lemma 32. *Assume $d \ge 3$ and $t \in (0,1)$. Let p_1 be a smooth density on S^d such that*

$$0 < m_1 \le p_1(z) \le M_1 < \infty.$$

Then for every $x \in S^d$,

$$\Big|\frac{d}{dt}\operatorname{Div} v(t,x)\Big| \le \frac{256\pi^2(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}.$$

Replacing the conditional expectation bounds by expectation bounds in the proof (based on 24), we obtain

$$\mathbb{E}\Big[\Big|\frac{d}{dt}\operatorname{Div} v(t,x)\Big|\Big] \le \frac{256\pi^2(d-1)^2}{(1-t)^2}.$$

Proof. We start by computing the time derivative of divergence. Recall

$$\operatorname{Div} v(t,x) = \frac{1}{1-t}\mathbb{E}[\operatorname{Div}_x \operatorname{Log}_x(X_1) \mid X_t = x] + \frac{1}{1-t}\mathbb{E}[\langle \operatorname{Grad}_x \log p_t(X_1 \mid x), \operatorname{Log}_x(X_1)\rangle \mid X_t = x].$$

Let $F(t, x_1)$ be smooth in t and integrable under $p_t(\cdot \mid x)$. Then

$$\frac{d}{dt}\mathbb{E}[F(t, X_1) \mid X_t = x] = \mathbb{E}[\partial_t F(t, X_1) \mid X_t = x] + \mathbb{E}[F(t, X_1)\, \partial_t \log p_t(X_1 \mid x) \mid X_t = x].$$

Applying this with $F(t, x_1) = \operatorname{Div}_x \operatorname{Log}_x(x_1)$ and $F(t, x_1) = \langle \operatorname{Grad}_x \log p_t(x_1 \mid x), \operatorname{Log}_x(x_1)\rangle$ respectively,

$$\begin{aligned}
&\frac{d}{dt}\mathbb{E}[\operatorname{Div}_x \operatorname{Log}_x(X_1) \mid X_t = x] = \mathbb{E}[\operatorname{Div}_x \operatorname{Log}_x(X_1)\, \partial_t \log p_t(X_1 \mid x) \mid X_t = x],\\
&\frac{d}{dt}\mathbb{E}[\langle \operatorname{Grad}_x \log p_t(X_1 \mid x), \operatorname{Log}_x(X_1)\rangle \mid X_t = x] = \mathbb{E}[\partial_t \langle \operatorname{Grad}_x \log p_t(X_1 \mid x), \operatorname{Log}_x(X_1)\rangle \mid X_t = x]\\
&\qquad + \mathbb{E}[\langle \operatorname{Grad}_x \log p_t(X_1 \mid x), \operatorname{Log}_x(X_1)\rangle\, \partial_t \log p_t(X_1 \mid x) \mid X_t = x].
\end{aligned}$$

Therefore, for fixed x,

$$\begin{aligned}
\partial_t \operatorname{Div} v(t, x) = &\frac{1}{(1-t)^2}\mathbb{E}\big[\operatorname{Div}_x \operatorname{Log}_x(X_1) + \langle \operatorname{Grad}_x \log p_t(X_1 \mid x), \operatorname{Log}_x(X_1)\rangle \mid X_t = x\big]\\
&+ \frac{1}{1-t}\mathbb{E}[\operatorname{Div}_x \operatorname{Log}_x(X_1)\, \partial_t \log p_t(X_1 \mid x) \mid X_t = x]\\
&+ \frac{1}{1-t}\mathbb{E}[\partial_t \langle \operatorname{Grad}_x \log p_t(X_1 \mid x), \operatorname{Log}_x(X_1)\rangle \mid X_t = x]\\
&+ \frac{1}{1-t}\mathbb{E}[\langle \operatorname{Grad}_x \log p_t(X_1 \mid x), \operatorname{Log}_x(X_1)\rangle\, \partial_t \log p_t(X_1 \mid x) \mid X_t = x].
\end{aligned}$$

Define

$$\begin{aligned}
T_1 &= \frac{1}{(1-t)^2}\mathbb{E}\big[|\operatorname{Div}_x \operatorname{Log}_x(X_1) + \langle \operatorname{Grad}_x \log p_t(X_1 \mid x), \operatorname{Log}_x(X_1)\rangle| \mid X_t = x\big],\\
T_2 &= \frac{1}{1-t}\mathbb{E}[|\operatorname{Div}_x \operatorname{Log}_x(X_1)\, \partial_t \log p_t(X_1 \mid x)| \mid X_t = x],\\
T_3 &= \frac{1}{1-t}\mathbb{E}[|\partial_t \langle \operatorname{Grad}_x \log p_t(X_1 \mid x), \operatorname{Log}_x(X_1)\rangle| \mid X_t = x],\\
T_4 &= \frac{1}{1-t}\mathbb{E}[|\langle \operatorname{Grad}_x \log p_t(X_1 \mid x), \operatorname{Log}_x(X_1)\rangle\, \partial_t \log p_t(X_1 \mid x)| \mid X_t = x].
\end{aligned}$$

For T_1, recall

$$\frac{1}{(1-t)^2}\mathbb{E}\big[\operatorname{Div}_x \operatorname{Log}_x(X_1) \mid X_t = x\big] \le \pi \frac{(d-1)}{(1-t)^2}\frac{4M_1}{m_1},$$

and

$$\begin{aligned}
&\frac{1}{(1-t)^2}\mathbb{E}\big[\langle \operatorname{Grad}_x \log p_t(X_1 \mid x), \operatorname{Log}_x(X_1)\rangle \mid X_t = x\big]\\
\le &\frac{1}{(1-t)^2}(1-t)\pi\mathbb{E}\big[\|\operatorname{Grad}_x \log p_t(X_1 \mid x)\| \mid X_t = x\big] \le \frac{8\pi(d-1)}{(1-t)^2}\frac{M_1}{m_1}.
\end{aligned}$$

Hence

$$T_1 \le \pi \frac{(d-1)}{(1-t)^2}\frac{4M_1}{m_1} + \frac{8\pi(d-1)}{(1-t)^2}\frac{M_1}{m_1} \le \frac{16\pi(d-1)}{(1-t)^2}\frac{M_1}{m_1}.$$

For T_2, recall

$$\mathbb{E}[|\operatorname{Div}_x \operatorname{Log}_x(X_1)|^2 \mid X_t = x] \le (d-1)^2 \frac{8\pi^2 M_1}{m_1}.$$

Hence we have

$$T_2 = \frac{1}{1-t}\mathbb{E}[|\operatorname{Div}_x \operatorname{Log}_x(X_1)\, \partial_t \log p_t(X_1 \mid x)| \mid X_t = x]$$

$$\leq \frac{1}{1-t}\Big(\mathbb{E}[|\operatorname{Div}_x \operatorname{Log}_x(X_1)|^2 \mid X_t = x]\mathbb{E}[|\partial_t \log p_t(X_1 \mid x)|^2 \mid X_t = x]\Big)^{\frac{1}{2}}$$
$$\leq \frac{1}{1-t}\Big((d-1)^2\frac{8\pi^2 M_1}{m_1}\frac{32\pi^2(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}\Big)^{\frac{1}{2}} \leq \frac{16\pi^2(d-1)^2}{(1-t)^2}\frac{M_1}{m_1},$$

where by Lemma 26,

$$\mathbb{E}\Big[\Big|\frac{\partial}{\partial t}\log p_t(X_1 \mid x)\Big|^2 \mid X_t = x\Big] \leq \frac{8}{(1-t)^2} + \frac{16\pi^2(d-1)^2}{(1-t)^2}\frac{M_1}{m_1} \leq \frac{32\pi^2(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}.$$

For T_3, since $\operatorname{Log}_x(x_1)$ does not depend on t,

$$|\partial_t \langle \operatorname{Grad}_x \log p_t(x_1 \mid x), \operatorname{Log}_x(x_1)\rangle|$$
$$\leq \|\partial_t \operatorname{Grad}_x \log p_t(x_1 \mid x)\| \, \|\operatorname{Log}_x(x_1)\| \leq (1-t)\pi \, \|\partial_t \operatorname{Grad}_x \log p_t(x_1 \mid x)\|.$$

Using Lemma 28, we have

$$\mathbb{E}[\|\partial_t \operatorname{Grad}_x \log p_t(X_1 \mid x)\| \mid X_t = x] \leq \frac{16\pi^2(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}.$$

Hence

$$\mathbb{E}[|\partial_t \langle \operatorname{Grad}_x \log p_t(X_1 \mid x), \operatorname{Log}_x(X_1)\rangle| \mid X_t = x] \leq (1-t)\pi \cdot \frac{16\pi^2(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}$$
$$= \frac{16\pi^3(d-1)^2}{(1-t)}\frac{M_1}{m_1}.$$

Last, we bound T_4. Similarly, by Cauchy-Schwarz and the bound we did for T_3,

$$\Big|\mathbb{E}[\langle \operatorname{Grad}_x \log p_t(X_1 \mid x), \operatorname{Log}_x(X_1)\rangle \, \partial_t \log p_t(X_1 \mid x) \mid X_t = x]\Big|$$
$$\leq \Big(\mathbb{E}[|\langle \operatorname{Grad}_x \log p_t(X_1 \mid x), \operatorname{Log}_x(X_1)\rangle|^2 \mid X_t = x]\Big)^{1/2}\Big(\mathbb{E}[|\partial_t \log p_t(X_1 \mid x)|^2 \mid X_t = x]\Big)^{1/2}.$$

We bound the first factor by

$$\mathbb{E}[|\langle \operatorname{Grad}_x \log p_t(X_1 \mid x), \operatorname{Log}_x(X_1)\rangle|^2 \mid X_t = x]$$
$$\leq (1-t)^2\pi^2\mathbb{E}[\|\operatorname{Grad}_x \log p_t(X_1 \mid x)\|^2 \mid X_t = x] \leq 32\pi^2(d-1)^2\frac{M_1}{m_1}.$$

Therefore

$$T_4 = \frac{1}{1-t}\Big|\mathbb{E}[\langle \operatorname{Grad}_x \log p_t(X_1 \mid x), \operatorname{Log}_x(X_1)\rangle \, \partial_t \log p_t(X_1 \mid x) \mid X_t = x]\Big|$$
$$\leq 32\pi^2\frac{(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}.$$

Together,

$$|\partial_t \operatorname{Div} v(t,x)| \leq T_1 + T_2 + T_3 + T_4$$
$$\leq \frac{16\pi(d-1)}{(1-t)^2}\frac{M_1}{m_1} + \frac{16\pi^2(d-1)^2}{(1-t)^2}\frac{M_1}{m_1} + \frac{16\pi^3(d-1)^2}{(1-t)^2}\frac{M_1}{m_1} + 32\pi^2\frac{(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}$$
$$\leq \frac{128\pi^2(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}.$$

Together with Lemma 31, we obtain

$$|\frac{d}{dt}\operatorname{Div} v(t,x)| = |\partial_t \operatorname{Div} v(t,x) + \langle \operatorname{Grad}_x \operatorname{Div} v(t,x), v(t,x)\rangle| \leq \frac{256\pi^2(d-1)^2}{(1-t)^2}\frac{M_1}{m_1}.$$

□

D.4 Regularity of $v(t,x)$ and $\log p_t$

Finally, we bound $\|v(t,x)\|$ and $\mathbb{E}[\|\operatorname{Grad}\log p_t(X_t)\|^2]$ in Lemma 33 and 34.

Lemma 33. *Assume $d \geq 2$ and $t \in (0,1)$. Let p_1 be a smooth density on S^d such that*

$$0 < m_1 \leq p_1(z) \leq M_1 < \infty.$$

For every $x \in S^d$,

$$\|v(t,x)\| \leq \pi.$$

Proof. Fix $x \in S^d$. By Jensen and the definition of v,

$$\|v(t,x)\| = \frac{1}{1-t}\Big\|\int_{S^d} \operatorname{Log}_x(x_1)\, p_t(x_1 \mid x)\, dV_g(x_1)\Big\| \leq \frac{1}{1-t}\int_{S^d} \|\operatorname{Log}_x(x_1)\|\, p_t(x_1 \mid x)\, dV_g(x_1).$$

On the support of $p_t(\cdot \mid x)$ we have $d(x,x_1) < (1-t)\pi$, hence

$$\|\operatorname{Log}_x(x_1)\| = d(x,x_1) \leq (1-t)\pi.$$

Therefore

$$\|v(t,x)\| \leq \frac{1}{1-t}\int_{S^d} (1-t)\pi\, p_t(x_1 \mid x)\, dV_g(x_1) = \pi \int_{S^d} p_t(x_1 \mid x)\, dV_g(x_1) = \pi,$$

since $p_t(\cdot \mid x)$ is a probability density. □

Lemma 34. *Assume $d \geq 3$ and $t \in (0,1)$. Let $p_0 \equiv \operatorname{Vol}(S^d)^{-1}$ be the uniform prior on S^d, and let p_1 be a smooth density on S^d such that*

$$0 < m_1 \leq p_1(z) \leq M_1 < \infty.$$

Let $X_0 \sim p_0$ and $X_1 \sim p_1$ be independent, and define the geodesic interpolation

$$X_t = \mathit{Exp}_{X_0}(t\, \mathit{Log}_{X_0}(X_1)).$$

Let p_t denote the marginal density of X_t with respect to dV_g. Then

$$\mathbb{E}\big[\|\mathit{Grad}\log p_t(X_t)\|^2\big] \leq \frac{8(d-1)^2}{(1-t)^2}.$$

Proof. Using Lemma 21, notice that $\log p_t(x) = \log Z_t(x) - \log \operatorname{Vol}(S^d)$, we have

$$\operatorname{Grad}\log p_t(x) = \operatorname{Grad}\log Z_t(x).$$

Take gradient, we obtain

$$\operatorname{Grad}\, Z_t(x) = \int_{S^d} p_1(x_1)\, \operatorname{Grad}_x J_t(x \mid x_1)\, dV_g(x_1).$$

Using $\operatorname{Grad}_x J_t = J_t \operatorname{Grad}_x \log J_t$ and the definition of the conditional density

$$p_t(x_1 \mid x) = \frac{p_1(x_1) J_t(x \mid x_1)}{Z_t(x)},$$

we obtain

$$\begin{aligned}\operatorname{Grad}\log Z_t(x) = \frac{\operatorname{Grad}\, Z_t(x)}{Z_t(x)} &= \int_{S^d} \operatorname{Grad}_x \log J_t(x \mid x_1)\, p_t(x_1 \mid x)\, dV_g(x_1) \\ &= \mathbb{E}\big[\operatorname{Grad}_x \log J_t(x \mid X_1) \mid X_t = x\big].\end{aligned}$$

Now we have

$$\operatorname{Grad}\log p_t(X_t) = \mathbb{E}[\operatorname{Grad}_x \log J_t(X_t \mid X_1) \mid X_t],$$

which implies

$$\|\operatorname{Grad}\log p_t(X_t)\|^2 \leq \mathbb{E}[\|\operatorname{Grad}_x \log J_t(X_t \mid X_1)\|^2 \mid X_t].$$

Taking expectation again gives

$$\mathbb{E}[\|\operatorname{Grad}\log p_t(X_t)\|^2] \leq \mathbb{E}\Big[\mathbb{E}[\|\operatorname{Grad}_x \log J_t(X_t \mid X_1)\|^2 \mid X_t]\Big] \leq \frac{8(d-1)^2}{(1-t)^2}.$$

where the last inequality follows from Lemma 25 and Lemma 24.

□

D.5 Finiteness of Score Regularity: Proof of Proposition 6

Proof. [Proof of Proposition 6] Recall that we assumed v is C^2 in (t, x). For $0 \le s \le r < 1$, let

$$\Phi_{s,r} : M \to M$$

denote the flow generated by v, that is, for each $x \in M$, the curve $\Phi_{s,u}(x)$ is the unique solution of

$$\frac{d}{du}\Phi_{s,u}(x) = v\big(u, \Phi_{s,u}(x)\big), \qquad \Phi_{s,s}(x) = x.$$

Since M is compact and v is of class C^2 in (t, x), for every $0 \le s, r < 1$, the map $\Phi_{s,r}$ is a C^2 diffeomorphism of M, and its inverse is $\Phi_{r,s}$. Define

$$q_r(x) := \frac{1}{\mathrm{Vol}_g(M)} \exp\left(-\int_0^r \text{Div } v\big(u, \Phi_{r,u}(x)\big)\, du\right).$$

We first show that the family $(q_r)_{0\le r<1}$ satisfies the continuity equation

$$\partial_r q_r + \text{Div}\left(q_r v(r, \cdot)\right) = 0.$$

Fix $x \in M$ representing the initial value at time being zero. Using the semigroup identity

$$\Phi_{u,r}\big(\Phi_{0,u}(x)\big) = \Phi_{0,u}(x) \qquad \text{for all } 0 \le r \le u < 1,$$

we obtain

$$\begin{aligned} q_r\big(\Phi_{0,r}(x)\big) &= \frac{1}{\mathrm{Vol}_g(M)} \exp\left(-\int_0^r \text{Div } v\big(u, \Phi_{r,u}(\Phi_{0,r}(x))\big)\, du\right) \\ &= \frac{1}{\mathrm{Vol}_g(M)} \exp\left(-\int_0^r \text{Div } v\big(u, \Phi_{0,u}(x)\big)\, du\right). \end{aligned}$$

Differentiating with respect to r gives

$$\frac{d}{dr} q_r\big(\Phi_{0,r}(x)\big) = -q_r\big(\Phi_{0,r}(x)\big)\, \text{Div } v\big(r, \Phi_{0,r}(x)\big).$$

On the other hand, by the chain rule,

$$\begin{aligned} \frac{d}{dr} q_r\big(\Phi_{0,r}(x)\big) &= \partial_r q_r\big(\Phi_{0,r}(x)\big) + \left\langle \mathrm{Grad}\, q_r\big(\Phi_{0,r}(x)\big), \frac{d}{dr}\Phi_{0,r}(x) \right\rangle \\ &= \partial_r q_r\big(\Phi_{0,r}(x)\big) + \left\langle \mathrm{Grad}\, q_r\big(\Phi_{0,r}(x)\big), v\big(r, \Phi_{0,r}(x)\big) \right\rangle. \end{aligned}$$

Combining the two identities yields

$$\partial_r q_r\big(\Phi_{0,r}(x)\big) + \left\langle \mathrm{Grad}\, q_r\big(\Phi_{0,r}(x)\big), v\big(r, \Phi_{0,r}(x)\big)\right\rangle + q_r\big(\Phi_{0,r}(x)\big)\, \text{Div } v\big(r, \Phi_{0,r}(x)\big) = 0.$$

Since $x \in M$ is arbitrary and $\Phi_{0,r}$ is onto, this implies

$$\partial_r q_r(y) + \langle \mathrm{Grad}\, q_r(y), v(r, y)\rangle + q_r(y)\, \text{Div } v(r, y) = 0 \qquad \text{for all } y \in M,\ 0 \le r < 1.$$

Recall the product rule of divergence $\text{Div}\left(q_r v(r, \cdot)\right) = \langle \mathrm{Grad}\, q_r, v(r, \cdot)\rangle + q_r\, \text{Div } v(r, \cdot)$, we conclude that

$$\partial_r q_r + \text{Div}\left(q_r v(r, \cdot)\right) = 0 \qquad \text{on } [0, 1) \times M.$$

Also, the initial values are the same: $q_0(x) = \frac{1}{\mathrm{Vol}_g(M)} = p_0(x)$. So we know, for every $t \in [0, 1)$,

$$p_t(x) = \frac{1}{\mathrm{Vol}_g(M)} \exp\left(-\int_0^t \text{Div } v\big(u, \Phi_{t,u}(x)\big)\, du\right) \qquad \text{for all } x \in M.$$

We now verify that $x \mapsto p_t(x)$ is of class C^1 for every fixed $t \in [0, 1)$. Since v is of class C^2 in (r, x), the map $(r, z) \mapsto \text{Div } v(r, z)$ is of class C^1 on $[0, t] \times M$. Therefore, for every fixed $u \in [0, t]$, the map

$$x \mapsto \text{Div } v\big(u, \Phi_{t,u}(x)\big)$$

is of class C^1 on M. Hence the map

$$x \mapsto \int_0^t \text{Div } v\big(u, \Phi_{t,u}(x)\big)\, du$$

is of class C^1 on M. It follows that $x \mapsto p_t(x)$ is of class C^1 on M.

Also, for every $x \in M$,

$$p_t(x) = \frac{1}{\text{Vol}_g(M)} \exp\left(-\int_0^t \text{Div } v\big(u, \Phi_{t,u}(x)\big)\, du\right) > 0.$$

Therefore $x \mapsto \log p_t(x)$ is of class C^1 on M, and hence

$$x \mapsto \text{Grad} \log p_t(x)$$

is continuous on M.

Finally, define

$$L_t^{\text{score}} := \sup_{x \in M} \|\text{Grad} \log p_t(x)\|^2 .$$

Since M is compact and $x \mapsto \text{Grad} \log p_t(x)$ is continuous, the above supremum is finite. Therefore

$$L_t^{\text{score}} < \infty.$$

Moreover,

$$\begin{aligned}\mathbb{E}\left[\|\text{Grad} \log p_t(X_t)\|^2\right] &= \int_M \|\text{Grad} \log p_t(x)\|^2 \, p_t(x)\, dV_g(x) \\ &\le \sup_{x \in M} \|\text{Grad} \log p_t(x)\|^2 = L_t^{\text{score}}.\end{aligned}$$

This proves the desired finiteness of L_t^{score}. □

E Hadamard Manifold Regularity Results

In this section, we work on Hadamard manifolds that is also a symmetric space ($\nabla R \equiv 0$), which covers SPD manifold with the affine invariant metric, denoted as $\text{SPD}\,(n)$.

We briefly discuss the similarity and difference between the regularity analysis in this section and that of the hypersphere.

- On a hypersphere, there exist cut points, resulting indicator function in the conditional density. We provide the formula for the conditional density function, on a general Hadamard manifold, see Lemma 35. We see that with non-positive curvature, there would be no cut points. Consequently, there would be no indicator function in the expression of the conditional density, guaranteeing better smoothness.
- The hypersphere is a compact manifold, but Hadamard manifold is non-compact. On a compact manifold, the vector field $v(t, x)$ itself is uniformly bounded and we can obtain pointwise regularity. But on a non-compact manifold, $\|v(t, x)\|$ could possibly blow up, and extra assumption would be needed to guarantee regularity in expectation. But we can still expect its norm to be bounded, in expectation.
- Moreover, from a high-level idea, the procedure for providing upper bounds on derivatives of $v(t, x)$ remains the same. For example, in Lemma 31, we expanded the gradient of $\text{Div } v(t, x)$ as terms involving $\mathbb{E}[\|\text{Grad}_x \text{Div}_x \text{Log}_x(X_1)\| \mid X_t = x]$, $\mathbb{E}[\|\text{Grad}_x \log p_t(X_1 \mid x)\|^2 \mid X_t = x]$, just to name a few. To establish regularity on Hadamard manifolds, we still have roughly the same expansion, involving the same collection of terms. The difference is that, instead of obtain uniform upper bounds (as in Lemma 25, 26, 27, and 28), our bounds will depend on radial distance function $r(x) = d(x, x_0)$ for some x_0.

Throughout this section, we use the following notation: $\kappa := \sqrt{-K_{\min}} > 0$ and the model function is defined as

$$s_{K_{\min}}(r) := \frac{1}{\kappa}\sinh(\kappa r), \qquad r \geq 0.$$

Define the *radial contraction map* and its inverse by

$$\Phi_{t,x_1}(x) := \operatorname{Exp}_{x_1}\big((1-t)\operatorname{Log}_{x_1}(x)\big),$$
$$\Psi_{t,x_1}(x) := \operatorname{Exp}_{x_1}\Big(\frac{1}{1-t}\operatorname{Log}_{x_1}(x)\Big).$$

and note that under our geodesic interpolation, $x_t = \Phi_{t,x_1}(x_0)$, and $x_0 = \Psi_{t,x_1}(x)$.

We first provide the density formula.

Lemma 35 (Density of $p_t(x_1 \mid x)$ on a Hadamard manifold)**.** *Let (M, g) be a complete, simply-connected, d-dimensional Riemannian manifold with non-positive curvature. Let p_0, p_1 be the prior and target distributions, assuming independence. Then the conditional density of X_1 given $X_t = x$ is*

$$p_t(x_1 \mid x) = \frac{p_1(x_1)\, p_0\big(\Psi_{t,x_1}(x)\big)\, J_t(x \mid x_1)}{\displaystyle\int_M p_1(z)\, p_0\big(\Psi_{t,z}(x)\big)\, J_t(x \mid z)\, dV_g(z)},$$
$$\textit{where} \quad J_t(x \mid x_1) = (1-t)^{-d}\, \frac{\big|\det(d\,\mathit{Exp}_{x_1})_{\frac{1}{1-t}\mathit{Log}_{x_1}(x)}\big|}{\big|\det(d\,\mathit{Exp}_{x_1})_{\mathit{Log}_{x_1}(x)}\big|}.$$

Furthermore, we have

$$p_t(x) = \int_M p_1(z)\, p_0\big(\Psi_{t,z}(x)\big)\, J_t(x \mid z)\, dV_g(z).$$

Proof.

Notice that Φ_{t,x_1} is a global diffeomorphism, and Ψ_{t,x_1} is its inverse. For fixed x_1 and x, the change-of-variables formula gives

$$dV_g(\Psi_{t,x_1}(x)) = \big|\det(d\Psi_{t,x_1})_x\big|\, dV_g(x),$$

which implies

$$p(\Psi_{t,x_1}(x), x_1)\, dV_g(\Psi_{t,x_1}(x))\, dV_g(x_1) = p_0\big(\Psi_{t,x_1}(x)\big)\, p_1(x_1)\, \big|\det(d\Psi_{t,x_1})_x\big|\, dV_g(x)\, dV_g(x_1).$$

We define

$$J_t(x \mid x_1) := \big|\det(d\Psi_{t,x_1})_x\big|.$$

Equivalently, the joint measure of (X_t, X_1) can be written as

$$p_t(x, x_1)\, dV_g(x)\, dV_g(x_1) = p_0\big(\Psi_{t,x_1}(x)\big)\, p_1(x_1)\, J_t(x \mid x_1)\, dV_g(x)\, dV_g(x_1),$$

Integrating the joint density over x_1 yields the marginal

$$p_t(x) = \int_M p_1(z)\, p_0\big(\Psi_{t,z}(x)\big)\, J_t(x \mid z)\, dV_g(z).$$

Therefore, by Bayes' rule,

$$p_t(x_1 \mid x) = \frac{p_t(x, x_1)}{p_t(x)} = \frac{p_1(x_1)\, p_0\big(\Psi_{t,x_1}(x)\big)\, J_t(x \mid x_1)}{\displaystyle\int_M p_1(z)\, p_0\big(\Psi_{t,z}(x)\big)\, J_t(x \mid z)\, dV_g(z)}.$$

It remains to compute $J_t(x \mid x_1)$. Recall that $\Psi_{t,x_1}(x) = \operatorname{Exp}_{x_1}\Big(\frac{1}{1-t}\operatorname{Log}_{x_1}(x)\Big)$. Using chain rule,

$$(d\Psi_{t,x_1})_x = (d\operatorname{Exp}_{x_1})_v \circ \big((1-t)^{-1} Id\big) \circ (d\operatorname{Log}_{x_1})_x.$$

Using $(d\operatorname{Log}_{x_1})_x = \big((d\operatorname{Exp}_{x_1})_{\operatorname{Log}_{x_1}(x)}\big)^{-1}$, we have

$$(d\Psi_{t,x_1})_x = (d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} \circ \big((1-t)^{-1} Id\big) \circ \big((d\operatorname{Exp}_{x_1})_{\operatorname{Log}_{x_1}(x)}\big)^{-1}.$$

Taking determinants,

$$J_t(x \mid x_1) = (d\Psi_{t,x_1})_x = (1-t)^{-d} \frac{\big|\det(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\big|}{\big|\det(d\operatorname{Exp}_{x_1})_{\operatorname{Log}_{x_1}(x)}\big|}.$$

□

Now we prove Proposition 14, which verifies Assumption 2.

Proof. [Proof of Proposition 14] The desired result is directly implied by Lemma 38 and Lemma 40:

$$\begin{aligned}
\mathbb{E}[\|v(t,x)\|^2] \lesssim& \mathbb{E}[d(x_0,x_1)^2] \lesssim dM_\lambda^{\frac{1}{2}},\\
\mathbb{E}[\|\nabla v(t,x)\|] \lesssim& \frac{d}{1-t} L_R \mathbb{E}[d(x_0,x_1)^2]^{\frac{1}{2}} \mathbb{E}[s_{K_{\min}}(d(x_0,x_1))^6]^{\frac{1}{2}}\\
\lesssim& \frac{d^{2+6\lambda}}{1-t} L_R M_\lambda^{\frac{1}{2}},\\
\mathbb{E}[|\frac{d}{dt} v(t,x)|] \lesssim& \frac{d}{1-t} L_R \mathbb{E}[d(x_0,x_1)^2 s_{K_{\min}}(d(x_0,x_1))^6]^{\frac{1}{2}} \mathbb{E}[d(x_0,x_1)^2]^{\frac{1}{2}}\\
\lesssim& \frac{d^{2+6\lambda}}{1-t} L_R M_\lambda^{\frac{1}{2}},\\
\mathbb{E}[\|\operatorname{Grad}_x \operatorname{Div} v(t,x)\|] \lesssim& \mathbb{E}[d(x_0,x_1) s_{K_{\min}}(d(x_0,x_1))^6] \frac{d^2}{(1-t)^2} L_R^3 \lesssim \frac{d^{3+12\lambda}}{(1-t)^2} L_R^3 M_\lambda^{\frac{1}{2}},\\
\mathbb{E}[|\frac{d}{dt} \operatorname{Div} v(t,x)|] \lesssim& \mathbb{E}[d(x_0,x_1)^2]^{\frac{1}{2}} \frac{d^2}{(1-t)^2} L_R^3 \mathbb{E}[d(x_1,x_0)^2 s_{K_{\min}}(d(x_0,x_1))^{12}]^{\frac{1}{2}}\\
\lesssim& \frac{d^{3+12\lambda}}{(1-t)^2} L_R^3 M_\lambda^{\frac{1}{2}},
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[\|\nabla v(t,x)\|^2] \lesssim& \frac{d^2}{(1-t)^2} L_R^2 \mathbb{E}[d(x_0,x_1)^4]^{\frac{1}{2}} \mathbb{E}[s_{K_{\min}}(d(x_0,x_1))^{12}]^{\frac{1}{2}} \lesssim \frac{d^{3+12\lambda}}{(1-t)^2} L_R^2 M_\lambda^{\frac{1}{2}},\\
\mathbb{E}[|\frac{d}{dt} v(t,x)|^2] \lesssim& \frac{d^2}{(1-t)^2} L_R^2 \mathbb{E}[d(x_0,x_1)^4 s_{K_{\min}}(d(x_0,x_1))^{12}]^{\frac{1}{2}} \mathbb{E}[d(x_0,x_1)^4]^{\frac{1}{2}} \lesssim \frac{d^{3+12\lambda}}{(1-t)^2} L_R^2 M_\lambda^{\frac{1}{2}},\\
\mathbb{E}[\|\operatorname{Grad}_x \operatorname{Div} v(t,x)\|^2] \lesssim& \mathbb{E}[d(x_0,x_1)^2 s_{K_{\min}}(d(x_0,x_1))^{12}] \frac{d^4}{(1-t)^4} L_R^6 \lesssim \frac{d^{5+24\lambda}}{(1-t)^4} L_R^6 M_\lambda.
\end{aligned}$$

□

E.1 Proof of Proposition 7

We start to prove the score regularity result. In the following Lemma, we decompose $\mathbb{E}\big[\|\operatorname{Grad} \log p_t(X_t)\|^2\big]$ as the sum of two terms.

Lemma 36. *For $X_t \sim p_t$,*

$$\mathbb{E}\big[\|\mathit{Grad}\, \log p_t(X_t)\|^2\big] \le 2\,\mathbb{E}\Big[\big\|\mathit{Grad}_x \log p_0(\Psi_{t,X_1}(X_t))\big\|^2\Big] + 2\,\mathbb{E}\Big[\big\|\mathit{Grad}_x \log J_t(X_t \mid X_1)\big\|^2\Big].$$

Proof. By Lemma 35,

$$p_t(x) = \int_M p_1(z)\, p_0\big(\Psi_{t,z}(x)\big)\, J_t(x \mid z)\, dV_g(z).$$

Taking the Riemannian gradient with respect to x and differentiating under the integral sign,

$$\text{Grad } p_t(x) = \int_M p_1(z)\ \text{Grad}_x\Big(p_0(\Psi_{t,z}(x))\, J_t(x \mid z)\Big)\, dV_g(z). \tag{16}$$

Using the identity $\text{Grad}\,(\phi) = \phi\ \text{Grad}\log\phi$, we have

$$\text{Grad}_x\Big(p_0(\Psi_{t,z}(x))\, J_t(x \mid z)\Big) = p_0(\Psi_{t,z}(x))\, J_t(x \mid z)\ \text{Grad}_x \log\Big(p_0(\Psi_{t,z}(x))\, J_t(x \mid z)\Big).$$

Plugging into (16) and dividing by $p_t(x)$ yields

$$\text{Grad}\log p_t(x) = \frac{\text{Grad}\ p_t(x)}{p_t(x)} = \int_M \text{Grad}_x \log\Big(p_0(\Psi_{t,z}(x))\, J_t(x \mid z)\Big)\, \frac{p_1(z)\, p_0(\Psi_{t,z}(x))\, J_t(x \mid z)}{p_t(x)}\, dV_g(z).$$

Observe that

$$\frac{p_1(z)\, p_0(\Psi_{t,z}(x))\, J_t(x \mid z)}{p_t(x)} = \frac{p_1(z)\, p_0(\Psi_{t,z}(x))\, J_t(x \mid z)}{\int_M p_1(z)\, p_0\big(\Psi_{t,z}(x)\big)\, J_t(x \mid z)\, dV_g(z)} = p_t(z \mid x).$$

Hence we have

$$\begin{aligned}\text{Grad}\log p_t(x) = \text{Grad}\log p_t(x) &= \int_M \text{Grad}_x \log\Big(p_0(\Psi_{t,z}(x))\, J_t(x \mid z)\Big)\, p_t(z \mid x)\, dV_g(z)\\ &= \mathbb{E}\Big[\,\text{Grad}_x \log\big(p_0(\Psi_{t,X_1}(x))J_t(x \mid X_1)\big) \mid X_t = x\Big].\end{aligned}$$

Next, apply Jensen's inequality,

$$\begin{aligned}\mathbb{E}[\|\,\text{Grad}\,\log p_t(x)\|^2] &= \mathbb{E}\Big[\|\mathbb{E}[\text{Grad}_x \log\big(p_0(\Psi_{t,X_1}(x))J_t(x \mid X_1)\big) \mid X_t = x]\|^2\Big]\\ &\le \mathbb{E}\Big[\mathbb{E}[\|\,\text{Grad}_x \log\big(p_0(\Psi_{t,X_1}(x))J_t(x \mid X_1)\big)\|^2 \mid X_t = x]\Big]\\ &= \mathbb{E}\Big[\|\,\text{Grad}_x \log\big(p_0(\Psi_{t,X_1}(x))J_t(x \mid X_1)\big)\|^2\Big].\end{aligned}$$

The result follows from

$$\text{Grad}_x \log\big(p_0(\Psi)\, J_t\big) = \text{Grad}_x \log p_0(\Psi) + \text{Grad}_x \log J_t, \qquad \|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2.$$

□

Now we prove Proposition 7.

Proof. [Proof of Proposition 7] From Lemma 36, we have

$$\mathbb{E}\big[\|\,\text{Grad}\,\log p_t(X_t)\|^2\big] \le 2\,\mathbb{E}\Big[\big\|\,\text{Grad}_x \log p_0(\Psi_{t,X_1}(X_t))\big\|^2\Big] + 2\,\mathbb{E}\Big[\big\|\,\text{Grad}_x \log J_t(X_t \mid X_1)\big\|^2\Big].$$

We show the following in Appendix E.3 (Lemma 41 and Lemma 42):

$$\begin{aligned}\|\,\text{Grad}_x \log J_t(x \mid x_1)\| &\le d\,\|(d\,\text{Log}_{x_1})_x\|\left(\frac{1}{1-t}\,\|\nabla(d\,\text{Exp}_{x_1})_{\frac{1}{1-t}\text{Log}_{x_1}(x)}\| + \|\nabla(d\,\text{Exp}_{x_1})_{\text{Log}_{x_1}(x)}\|\right),\\ \Big\|\,\text{Grad}_x \log p_0(\Psi_{t,X_1}(X_t))\Big\| &\le 2dd(x_0,z)\Big\|(d\,\text{Exp}_{x_1})_{\frac{1}{1-t}\text{Log}_{x_1}(x)}\Big\|_{\text{op}}\ \frac{1}{1-t}\ \big\|(d\,\text{Log}_{x_1})_x\big\|_{\text{op}}.\end{aligned}$$

We also prove the following in Lemma 44 and Lemma 46:

$$\begin{aligned}\|(d\,\text{Log}_{x_1})_x\| &\le 1,\\ \Big\|(d\,\text{Exp}_{x_1})_{\frac{1}{1-t}\text{Log}_{x_1}(x)}\Big\|_{\text{op}} &\le s_{K_{\min}}(\|\tfrac{1}{1-t}\,\text{Log}_{x_1}(x)\|)/\|\tfrac{1}{1-t}\,\text{Log}_{x_1}(x)\| = s_{K_{\min}}(d(x_0,x_1))/d(x_0,x_1),\\ \|\nabla(d\,\text{Exp}_{x_1})_{\frac{1}{1-t}\text{Log}_{x_1}(x)}\| &\le \frac{16}{3}s_{K_{\min}}(d(x_0,x_1)/2)^2 L_R s_{K_{\min}}(d(x_0,x_1)).\end{aligned}$$

So together,

$$\|\,\text{Grad}_x \log J_t(x \mid x_1)\| \lesssim \frac{d}{1-t} L_R s_{K_{\min}}(d(x_0,x_1))^3,$$

$$\Big\| \operatorname{Grad}_x \log p_0(\Psi_{t,X_1}(X_t)) \Big\| \lesssim \frac{d}{1-t} s_{K_{\min}}(d(x_0,x_1)) d(x_0,z).$$

Using Lemma 38, we have

$$\begin{aligned}\mathbb{E}\big[\| \operatorname{Grad} \log p_t(X_t)\|^2\big] &\lesssim \frac{d^2}{(1-t)^2} L_R^2 \mathbb{E}[s_{K_{\min}}(d(x_0,x_1))^6] \lesssim \frac{d^2}{(1-t)^2} L_R^2 \mathbb{E}[s_{K_{\min}}(d(x_0,x_1))^6] \\ &\lesssim \frac{d^{2+12\lambda}}{(1-t)^2} L_R^2 M_\lambda,\end{aligned}$$

where $\lambda = \max\{1, \kappa\}$.

□

E.2 Auxiliary Results for Regularity for Hadamard and SPD manifolds

E.2.1 Expectation Control

Due to the non-compact nature of a Hadamard manifold as well as curvature distortion, bounding expectations of the form

$$\mathbb{E}[d(X_0,X_1)^2 \, s_{K_{\min}}\big(d(X_0,X_1)\big)^a], \quad a \in \mathbb{N},$$

is a key step to establish regularity. We show that by choosing X_0 following a Riemannian Gaussian distribution, the above expectation can be controlled for a class of data distribution X_1 satisfying certain moment condition (5).

Lemma 37. *Let (M, g) be a complete, simply connected d-dimensional Riemannian manifold (Hadamard) and assume*

$$K_{\min} \le \sec \le 0, \qquad \textit{where } K_{\min} < 0 \textit{ is independent of } d.$$

Fix a basepoint $z \in M$ and set $r(x) := d(x,z)$. Define

$$p_0(x) = \frac{1}{Z} e^{(-\beta\, r(x)^m)}, \qquad Z := \int_M e^{-\beta\, r(x)^m}\, dV_g(x), \qquad \textit{where} \qquad \beta = d, m = 2.$$

Let $\lambda \ge 0$ be independent of d. Then for $b \in \{0,1,2,3,4\}$,

$$\mathbb{E}_{p_0}\big[r(x)^b e^{\lambda r(x)}\big] = \mathcal{O}(d^{2\lambda}(\log d)^b).$$

Proof. We first write the expectation as an integral over tangent space. Since M is Hadamard, $\operatorname{Exp}_z : T_zM \to M$ is a global C^∞ diffeomorphism. We equip the vector space T_zM with the inner product g, and denote by dv the corresponding Lebesgue measure. For any measurable $F : M \to [0,\infty]$, the change-of-variables formula gives

$$\int_M F(x)\, dV_g(x) = \int_{T_zM} F(\operatorname{Exp}_z(v)) \, \big|\det\big(d(\operatorname{Exp}_z)_v\big)\big| \, dv.$$

We now apply this formula with $F_1(x) = e^{-d\, r(x)^2}$, $F_2(x) = r(x)^b e^{\lambda r(x)} e^{-d\, r(x)^2}$. Since $r(\operatorname{Exp}_z(v)) = \|v\|$, we obtain

$$\mathbb{E}_{p_0}[r^b e^{\lambda r}] = \frac{\int_{T_zM} \|v\|^b e^{\lambda\|v\|} e^{-d\|v\|^2} \, \big|\det(d(\operatorname{Exp}_z)_v)\big| \, dv}{\int_{T_zM} e^{-d\|v\|^2} \, \big|\det(d(\operatorname{Exp}_z)_v)\big| \, dv}. \tag{17}$$

Recall that using comparison theorem (note that the $d-1$ comes from $\det$) we have

$$1 \le \big|\det(d(\operatorname{Exp}_z)_{r\theta})\big| \le \Big(\frac{s_{K_{\min}}(r)}{r}\Big)^{d-1}, \qquad \forall r > 0. \tag{18}$$

Since $\frac{s_{K_{\min}}(r)}{r} = \frac{\sinh(\kappa r)}{\kappa r} \le e^{\kappa r}$, we have

$$\big|\det(d(\operatorname{Exp}_z)_{r\theta})\big| \le e^{\kappa(d-1)r}. \tag{19}$$

We now integrate in T_zM using Euclidean polar coordinates: $v = r\theta$ with $r \in [0, \infty)$, $\theta \in \mathbb{S}^{d-1}$, and

$$dv = r^{d-1}\, dr\, d\theta.$$

We split at $R = 2\log d$ and bound the expectation. Consider the decomposition below

$$\mathbb{E}_{p_0}[r^b e^{\lambda r}] = \mathbb{E}_{p_0}[r^b e^{\lambda r}\mathbf{1}_{\{r\le R\}}] + \mathbb{E}_{p_0}[r^b e^{\lambda r}\mathbf{1}_{\{r>R\}}].$$

We first consider the central part. On $\{r \le R\}$, $r^b e^{\lambda r} \le R^b e^{\lambda R}$, hence

$$\mathbb{E}_{p_0}[r^b e^{\lambda r}\mathbf{1}_{\{r\le R\}}] \le R^b e^{\lambda R} = (2\log d)^b\, d^{2\lambda}. \tag{20}$$

We next consider the tail part. Write the tail contribution as N_{tail}/Z. Using change of variable formula and the upper bound (19), we obtain

$$\begin{aligned} N_{\mathrm{tail}} &:= \int_{\{x:\, r(x)>R\}} r(x)^b e^{\lambda r(x)} e^{-dr(x)^2}\, dV_g(x) \\ &= \int_{\{v:\, \|v\|>R\}} \|v\|^b e^{\lambda\|v\|} e^{-d\|v\|^2} \big|\det(d(\mathrm{Exp}_z)_v)\big|\, dv \\ &\le |\mathbb{S}^{d-1}| \int_R^\infty r^{d-1+b} e^{(-dr^2+(\lambda+\kappa(d-1))r)}\, dr. \end{aligned} \tag{21}$$

Define

$$\Psi_d(r) := -dr^2 + (\lambda + \kappa(d-1))r + (d-1+b)\log r.$$

We justify that this $\Psi_d(r)$ is dominated by the $-dr^2$ term. For $r \ge R = 2\log d$, we have $(d-1+b)\log r \le \frac{d}{4}r^2, \forall d \ge 4$. Also, for $r \ge R$ and d large, $\kappa(d-1)r \le \frac{d}{4}r^2$ because this is equivalent to $r \ge 4\kappa(1-1/d)$, which holds once $R \ge 4\kappa$ (i.e. $d \ge e^{2\kappa}$). Finally, $\lambda r \le \frac{d}{8}r^2$ holds for all $r \ge R$ once $d \ge 8\lambda$. Summing these inequalities yields, there exists some constant d_1 s.t. for all $d \ge d_1$ and all $r \ge R$,

$$(\lambda + \kappa(d-1))r + (d-1+b)\log r \le \frac{d}{2}r^2, \quad \text{hence} \quad \Psi_d(r) \le -\frac{d}{2}r^2.$$

Therefore, for $d \ge d_1$,

$$\begin{aligned} \int_R^\infty e^{\Psi_d(r)}\, dr \le \int_R^\infty e^{-(d/2)r^2}\, dr &= \int_R^\infty \frac{1}{rd}\Big(-\frac{d}{dr}e^{-(d/2)r^2}\Big)\, dr \le \frac{1}{Rd}\int_R^\infty \Big(-\frac{d}{dr}e^{-(d/2)r^2}\Big)\, dr \\ &= \frac{1}{dR}\exp\Big(-\frac{d}{2}R^2\Big) = \frac{1}{dR}\exp\big(-2d(\log d)^2\big). \end{aligned}$$

Plugging into (21) gives

$$N_{\mathrm{tail}} \le |\mathbb{S}^{d-1}| \cdot \frac{1}{dR}\exp\big(-2d(\log d)^2\big). \tag{22}$$

For the denominator Z, recall that $\big|\det(d(\mathrm{Exp}_z)_v)\big| \ge 1$, so we get

$$\int_{T_zM} e^{-d\|v\|^2}\big|\det(d(\mathrm{Exp}_z)_v)\big|\, dv \ge \int_{T_zM} e^{-d\|v\|^2}\, dv = \Big(\frac{\pi}{d}\Big)^{d/2}.$$

Hence, using (22),

$$\mathbb{E}_{p_0}[r^b e^{\lambda r}\mathbf{1}_{\{r>R\}}] = \frac{N_{\mathrm{tail}}}{Z} \le \frac{|\mathbb{S}^{d-1}|}{dR}\exp\big(-2d(\log d)^2\big)\cdot\Big(\frac{d}{\pi}\Big)^{d/2} = \frac{|\mathbb{S}^{d-1}|}{2d\log d}d^{-2d\log d}\cdot\Big(\frac{d}{\pi}\Big)^{d/2} \lesssim 1.$$

Together, we conclude that for all $d \ge d_1$, combining (20) with the tail bound gives

$$\mathbb{E}_{p_0}[r^b e^{\lambda r}] \le (2\log d)^b d^{2\lambda} + 1 = \mathcal{O}(d^{2\lambda}(\log d)^b).$$

□

Now we present intermediate steps.

Lemma 38 (Auxiliary bounds for expectation of $s_{K_{\min}}$)**.** *Let M be a Hadamard manifold. Fix $z \in M$. Choose prior as $X_0 \sim p_0$ where*

$$p_0(x) \propto \exp\big(-d\, d(x,z)^2\big).$$

Assume Assumption 1 holds. Let $\kappa = \sqrt{-K_{\min}}$ and $\lambda_0 = a\max\{1,\kappa\}$, where $0 \le a \le a_0$ is some constant. Assume

$$\max\left\{\mathbb{E}[d(X_1,z)^4 e^{\lambda_0 d(X_1,z)}], \mathbb{E}[e^{\lambda_0 d(X_1,z)}]\right\} \le M.$$

Then we have, for $b \in \{0,1,2,3,4\}$,

$$\mathbb{E}\Big[d(X_0,X_1)^b\, s_{K_{\min}}\big(d(X_0,X_1)\big)^a\Big] \lesssim M\mathbb{E}_{X_0\sim p_0}\Big[d(X_0,z)^b e^{\lambda_0 d(X_0,z)}\Big] \lesssim d^{2\lambda_0}(\log d)^b M,$$
$$\mathbb{E}\Big[d(X_0,z)^b\, s_{K_{\min}}\big(d(X_0,X_1)\big)^a\Big] \lesssim M\mathbb{E}_{X_0\sim p_0}\Big[d(X_0,z)^b e^{\lambda_0 d(X_0,z)}\Big] \lesssim d^{2\lambda_0}(\log d)^b M.$$

Proof. Observe that

$$s_{K_{\min}}(d(x_0,x_1))^a = \frac{\sinh^a(\kappa d(x_0,x_1))}{\kappa^a} \le (e^{\max\{1,\kappa\}d(x_0,x_1)})^a.$$

Hence

$$d(x_0,x_1)^b\, s_{K_{\min}}\big(d(x_0,x_1)\big)^a \le d(x_0,x_1)^b (e^{\max\{1,\kappa\}d(x_0,x_1)})^a. \tag{23}$$

Taking expectation, it suffices to bound $\mathbb{E}\big[d(x_0,x_1)^b e^{\lambda_0 d(x_0,x_1)}\big]$ with $\lambda_0 = a\max\{1,\kappa\}$.

The triangle inequality gives

$$d(x_0,x_1) \le d(x_0,z) + d(x_1,z).$$

Therefore,

$$\begin{aligned} d(x_0,x_1)^b e^{\lambda_0 d(x_0,x_1)} \le& \big(d(x_0,z)+d(x_1,z)\big)^b \exp\big(\lambda_0(d(x_0,z)+d(x_1,z))\big) \\ \lesssim& (d(x_0,z)^b + d(x_1,z)^b)\exp(\lambda_0 d(x_0,z))\exp(\lambda_0 d(x_1,z)). \end{aligned}$$

Take expectation and use independence of X_0 and X_1:

$$\begin{aligned} \mathbb{E}\Big[d(X_0,X_1)^b e^{\lambda_0 d(X_0,X_1)}\Big] \le& 2\,\mathbb{E}\Big[e^{\lambda_0 d(X_1,z)}\Big]\,\mathbb{E}_{X_0\sim p_0}\Big[d(X_0,z)^b e^{\lambda_0 d(X_0,z)}\Big] \\ &+ 2\,\mathbb{E}\Big[d(X_1,z)^b e^{\lambda_0 d(X_1,z)}\Big]\,\mathbb{E}_{X_0\sim p_0}\Big[e^{\lambda_0 d(X_0,z)}\Big]. \end{aligned} \tag{24}$$

By Assumption, both $\mathbb{E}\Big[e^{\lambda_0 d(X_1,z)}\Big]$ and $\mathbb{E}\Big[d(X_1,z)^b e^{\lambda_0 d(X_1,z)}\Big]$ are controlled by M. It hence follows that

$$\mathbb{E}\Big[d(X_0,X_1)^b e^{\lambda_0 d(X_0,X_1)}\Big] \lesssim M\mathbb{E}_{X_0\sim p_0}\Big[d(X_0,z)^b e^{\lambda_0 d(X_0,z)}\Big] \lesssim d^{2\lambda_0}(\log d)^b M,$$

where we applied Lemma 37.

□

E.2.2 Regularity Control

The following lemma summarizes the building blocks needed to establish regularity, serving as the same purpose as Lemma 25, 26, 27, and 28.

Lemma 39. *On a Hadamard manifold that is also a symmetric space ($\nabla R \equiv 0$), with prior distribution being $p_0 \propto \exp(-dd(x_0,z)^2)$, assume Assumption 1, we have the following bounds.*

$$\begin{aligned} \|\mathit{Grad}_x \log J_t(x \mid x_1)\| \lesssim& d\,\frac{1}{1-t} L_R s_{K_{\min}}(d(x_0,x_1))^3, \\ \Big\|\mathit{Grad}_x \log p_0(\Psi_{t,x_1}(x))\Big\| \lesssim& d\frac{1}{1-t} s_{K_{\min}}(d(x_0,x_1)) d(x_0,z), \end{aligned}$$

$$
\begin{aligned}
|\partial_t \log J_t(x \mid x_1)| \lesssim& \frac{d}{1-t} d(x_0, x_1) s_{K_{\min}}(d(x_0, x_1))^3 L_R, \\
|\partial_t \log p_0(\Psi_{t,x_1}(x))| \lesssim& \frac{d}{1-t} d(x_0, z) s_{K_{\min}}(d(x_0, x_1)), \\
\|\nabla_x^2 \log p_0(\Psi_{t,x_1}(x))\|_{\mathrm{op}} \lesssim& \frac{d}{(1-t)^2} s_{K_{\min}}(d(x_0, x_1))^4 L_R d(x_0, z), \\
\|\nabla_x^2 \log J_t(x \mid x_1)\|_{\mathrm{op}} \lesssim& \frac{d}{(1-t)^2} L_R^3 s_{K_{\min}}(d(x_0, x_1))^6, \\
\left\| \partial_t \, \mathit{Grad}_x \log p_0(\Psi_{t,x_1}(x)) \right\| \lesssim& d(x_0, z) \frac{d}{(1-t)^2} d(x_0, x_1) L_R s_{K_{\min}}(d(x_0, x_1))^3, \\
\|\partial_t \, \mathit{Grad}_x \log J_t(x \mid x_1)\| \lesssim& \frac{d}{(1-t)^2} L_R^3 d(x_1, x_0) s_{K_{\min}}(d(x_0, x_1))^6, \\
\|\nabla_x \mathit{Log}_x(x_1)\|_{\mathrm{op}}^2 \lesssim& d(x_0, x_1)^2, \\
|\mathit{Div}_x \mathit{Log}_x(x_1)|^2 \lesssim& d^2 d(x_0, x_1)^2, \\
\|\mathit{Grad}\ \mathit{Div}_x \mathit{Log}_x(x_1)\| \lesssim& \frac{d L_R}{d(x, y)} s_{K_{\min}}(d(x, y))^2.
\end{aligned}
$$

On $\mathrm{SPD}(n)$*, the last inequality can be sharpened to* $\|\mathit{Grad}\ \mathit{Div}_x \mathit{Log}_x(x_1)\| \lesssim d d(x_0, x_1)^{\frac{3}{2}}$*. Here we emphasize that* $\mathit{Log}_x(x_1)$ *is viewed as a vector field, for fixed* x_1*, so* $\nabla \mathit{Log}_x(x_1)$ *is covariant derivative of the vector field* $\mathit{Log}_x(x_1)$*. Consequently,*

$$
\begin{aligned}
\mathbb{E}[\|\mathit{Grad}_x \log p_t(x_1 \mid x)\|^2 \mid X_t = x] \lesssim& \frac{d^2}{(1-t)^2} L_R^2 \mathbb{E}[s_{K_{\min}}(d(x_0, x_1))^6], \\
\mathbb{E}[\|\mathit{Grad}_x \log p_t(x_1 \mid x)\|^4 \mid X_t = x] \lesssim& \frac{d^4}{(1-t)^4} L_R^4 \mathbb{E}[s_{K_{\min}}(d(x_0, x_1))^{12}], \\
\mathbb{E}[|\partial_t \log p_t(x_1 \mid x)|^2 \mid X_t = x] \lesssim& \frac{d^2}{(1-t)^2} L_R^2 \mathbb{E}[d(x_0, x_1)^2 s_{K_{\min}}(d(x_0, x_1))^6], \\
\mathbb{E}[|\partial_t \log p_t(x_1 \mid x)|^4 \mid X_t = x] \lesssim& \frac{d^4}{(1-t)^4} L_R^4 \mathbb{E}[d(x_0, x_1)^4 s_{K_{\min}}(d(x_0, x_1))^{12}], \\
\mathbb{E}[\|\nabla_x^2 \log p_t(x_1 \mid x)\|_{\mathrm{op}}^2 \mid X_t = x] \lesssim& \frac{d^2}{(1-t)^4} L_R^6 \mathbb{E}[s_{K_{\min}}(d(x_0, x_1))^{12}] \\
&+ \frac{d^4}{(1-t)^4} L_R^4 \mathbb{E}[s_{K_{\min}}(d(x_0, x_1))^6]^2, \\
\mathbb{E}[\|\nabla_x^2 \log p_t(x_1 \mid x)\|_{\mathrm{op}}^4 \mid X_t = x] \lesssim& \frac{d^4}{(1-t)^8} L_R^{12} \mathbb{E}[s_{K_{\min}}(d(x_0, x_1))^{24}] \\
&+ \frac{d^8}{(1-t)^8} L_R^8 \mathbb{E}[s_{K_{\min}}(d(x_0, x_1))^{12}]^2, \\
\mathbb{E}[\|\partial_t \, \mathit{Grad}_x \log p_t(x_1 \mid x)\|^2 \mid X_t = x] \lesssim& \frac{d^2}{(1-t)^4} L_R^6 \mathbb{E}[d(x_1, x_0)^2 s_{K_{\min}}(d(x_0, x_1))^{12}] \\
&+ \frac{d^4}{(1-t)^4} L_R^4 \mathbb{E}[s_{K_{\min}}(d(x_0, x_1))^6] \mathbb{E}[d(x_0, x_1)^2 s_{K_{\min}}(d(x_0, x_1))^6].
\end{aligned}
$$

Proof. Throughout the proof, we use $r(x)$ to denote the radial distance function $d(x, z)$, where z is the center of $p_0 \propto \exp(-\beta d(x_0, z)^2)$. We show the following in Appendix E.3 (Lemma 41 and Lemma 42):

$$
\begin{aligned}
\|\operatorname{Grad}_x \log J_t(x \mid x_1)\| \leq& d \, \|(d \operatorname{Log}_{x_1})_x\| \left(\frac{1}{1-t} \, \|\nabla (d \operatorname{Exp}_{x_1})_{\frac{1}{1-t} \operatorname{Log}_{x_1}(x)}\| + \|\nabla (d \operatorname{Exp}_{x_1})_{\operatorname{Log}_{x_1}(x)}\| \right), \\
\left\| \operatorname{Grad}_x \log p_0(\Psi_{t,X_1}(X_t)) \right\| \leq& 2 d d_g(x_0, z) \left\| (d \operatorname{Exp}_{x_1})_{\frac{1}{1-t} \operatorname{Log}_{x_1}(x)} \right\|_{\mathrm{op}} \, \frac{1}{1-t} \, \left\| (d \operatorname{Log}_{x_1})_x \right\|_{\mathrm{op}}.
\end{aligned}
$$

We also prove the following in (Lemma 44 and Lemma 46:

$$\begin{aligned}
\|(d\operatorname{Log}_{x_1})_x\| &\le 1,\\
\left\|\left(\nabla(d\operatorname{Log}_x)\right)_y\right\| &\le \left\|\left(\nabla(d\operatorname{Exp}_x)\right)_{\operatorname{Log}_x(y)}\right\|,\\
\left\|(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\right\|_{\mathrm{op}} &\le s_{K_{\min}}(\|\tfrac{1}{1-t}\operatorname{Log}_{x_1}(x)\|)/\|\tfrac{1}{1-t}\operatorname{Log}_{x_1}(x)\| \lesssim s_{K_{\min}}(d(x_0,x_1)),\\
\|\nabla(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\| &\le \frac{16}{3}s_{K_{\min}}(d(x_0,x_1)/2)^2 L_R s_{K_{\min}}(d(x_0,x_1)) \lesssim L_R s_{K_{\min}}(d(x_0,x_1))^3,\\
\|\nabla^2(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\| &\lesssim L_R^3 s_{K_{\min}}(d(x_0,x_1))^5.
\end{aligned}$$

We estimate the terms as follows. For gradient and time derivative, we have

$$\begin{aligned}
\|\operatorname{Grad}_x \log J_t(x \mid x_1)\| \lesssim& d\,\frac{1}{1-t}L_R s_{K_{\min}}(d(x_0,x_1))^3,\\
\left\|\operatorname{Grad}_x \log p_0(\Psi_{t,X_1}(X_t))\right\| \lesssim& d\frac{1}{1-t}s_{K_{\min}}(d(x_0,x_1))d(x_0,z),
\end{aligned}$$

and

$$\begin{aligned}
|\partial_t \log J_t(x \mid x_1)| \le& \frac{d}{1-t} + \frac{d}{(1-t)^2}\,\|\operatorname{Log}_{x_1}(x)\|\,\|\nabla(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\|\\
\lesssim& \frac{d}{1-t}d(x_0,x_1)s_{K_{\min}}(d(x_0,x_1))^3 L_R,\\
|\partial_t \log p_0(\Psi_{t,x_1}(x))| \le& 2\beta r\left\|(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\right\|_{\mathrm{op}}\,\frac{1}{(1-t)^2}\,\|\operatorname{Log}_{x_1}(x)\|\\
\lesssim& \frac{d}{1-t}d(x_0,z)s_{K_{\min}}(d(x_0,x_1)).
\end{aligned}$$

For second covariant derivative,

$$\begin{aligned}
&\|\nabla_x^2 \log p_0(\Psi_{t,x_1}(x))\|_{\mathrm{op}}\\
\le&\left(\left\|(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\right\|_{\mathrm{op}}\,\frac{1}{1-t}\,\left\|(d\operatorname{Log}_{x_1})_x\right\|_{\mathrm{op}}\right)^2\left(2\beta + 2\beta r\|\nabla^2 r\|_{\mathrm{op}}\Big|_{\Psi_{t,x_1}(x)}\right)\\
&+\left\|\nabla(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\right\|\left(\frac{1}{1-t}\left\|(d\operatorname{Log}_{x_1})_x\right\|_{\mathrm{op}}\right)^2 2\beta r\\
&+\left\|(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\right\|_{\mathrm{op}}\,\frac{1}{1-t}\left\|\nabla(d\operatorname{Log}_{x_1})_x\right\| 2\beta r\\
\lesssim&\frac{1}{(1-t)^2}s_{K_{\min}}(d(x_0,x_1))^2(dd(x_0,z)\kappa\coth(\kappa d(x_0,z)))\\
&+\frac{d}{(1-t)^2}s_{K_{\min}}(d(x_0,x_1))^3 L_R d(x_0,z) + \frac{d}{1-t}d(x_0,z)s_{K_{\min}}(d(x_0,x_1))^4 L_R\\
\lesssim&\frac{d}{(1-t)^2}s_{K_{\min}}(d(x_0,x_1))^4 L_R d(x_0,z),
\end{aligned}$$

where notice that $x\coth(x)$ is of order x. And we have

$$\begin{aligned}
&\|\nabla_x^2 \log J_t(x \mid x_1)\|_{\mathrm{op}}\\
\le& d\,\|(d\operatorname{Log}_{x_1})_x\|^2\left[\frac{1}{(1-t)^2}\left(\|\nabla(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\|^2 + \|\nabla^2(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\|\right)\right.\\
&\left.+\left(\|\nabla(d\operatorname{Exp}_{x_1})_{\operatorname{Log}_{x_1}(x)}\|^2 + \|\nabla^2(d\operatorname{Exp}_{x_1})_{\operatorname{Log}_{x_1}(x)}\|\right)\right]\\
&+ d\,\|\nabla(d\operatorname{Log}_{x_1})_x\|\left(\frac{1}{1-t}\,\|\nabla(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\| + \|\nabla(d\operatorname{Exp}_{x_1})_{\operatorname{Log}_{x_1}(x)}\|\right)
\end{aligned}$$

$$
\begin{aligned}
&\lesssim \frac{d}{(1-t)^2}(L_R^2 s_{K_{\min}}(d(x_0,x_1))^6 + L_R^3 s_{K_{\min}}(d(x_0,x_1))^5) + \frac{d}{1-t} L_R^2 s_{K_{\min}}(d(x_0,x_1))^6 \\
&\lesssim \frac{d}{(1-t)^2} L_R^3 s_{K_{\min}}(d(x_0,x_1))^6.
\end{aligned}
$$

For time derivative of gradient,

$$
\begin{aligned}
&\left\| \partial_t \operatorname{Grad}_x \log p_0(\Psi_{t,x_1}(x)) \right\| \\
\le& \Big(\Big\| (d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} \Big\|_{\mathrm{op}} \ \frac{1}{1-t} \ \big\| (d\operatorname{Log}_{x_1})_x \big\|_{\mathrm{op}} \Big) \times \Big(2\beta + 2\beta r \|\nabla^2 r\|_{\mathrm{op}} \Big|_{\Psi_{t,x_1}(x)} \Big) \\
&\times \Big\| (d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} \Big\|_{\mathrm{op}} \ \frac{1}{(1-t)^2} \ \| \operatorname{Log}_{x_1}(x) \| \\
&+ \Bigg[\frac{1}{(1-t)^2} \Big\| (d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} \Big\|_{\mathrm{op}} \ \big\| (d\operatorname{Log}_{x_1})_x \big\|_{\mathrm{op}} \\
&+ \frac{1}{(1-t)^3} \| \operatorname{Log}_{x_1}(x) \| \Big\| \nabla (d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} \Big\| \ \big\| (d\operatorname{Log}_{x_1})_x \big\|_{\mathrm{op}} \Bigg] 2\beta r \\
\lesssim& \Big(d + d d(x_0,z) \Big) s_{K_{\min}}(d(x_0,x_1))^2 \frac{1}{(1-t)^2} \\
&+ d d(x_0,z) \Big(\frac{1}{(1-t)^2} s_{K_{\min}}(d(x_0,x_1)) + \frac{1}{(1-t)^2} d(x_0,x_1) L_R s_{K_{\min}}(d(x_0,x_1))^3 \Big) \\
\lesssim& d(x_0,z) \frac{d}{(1-t)^2} d(x_0,x_1) L_R s_{K_{\min}}(d(x_0,x_1))^3.
\end{aligned}
$$

And we have

$$
\begin{aligned}
&\| \partial_t \operatorname{Grad}_x \log J_t(x \mid x_1) \| \\
\le& \| (d\operatorname{Log}_{x_1})_x \| \Bigg[\frac{d}{(1-t)^2} \| \nabla (d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} \| \\
&+ \frac{d}{(1-t)^3} \| \operatorname{Log}_{x_1}(x) \| \Big(\| \nabla (d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} \|^2 + \| \nabla^2 (d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} \| \Big) \Bigg] \\
\lesssim& \frac{d}{(1-t)^2} L_R s_{K_{\min}}(d(x_0,x_1))^3 + \frac{d}{(1-t)^2} d(x_1,x_0) \Big(L_R^2 s_{K_{\min}}(d(x_0,x_1))^6 + L_R^3 s_{K_{\min}}(d(x_0,x_1))^5 \Big) \\
\lesssim& \frac{d}{(1-t)^2} L_R^3 d(x_1,x_0) s_{K_{\min}}(d(x_0,x_1))^6.
\end{aligned}
$$

The three inequalities on Log follow from Appendix E.5:

$$
\begin{aligned}
\| \nabla_x \operatorname{Log}_x(y) \|_{\mathrm{op}}^2 &\le 2 + 2\, d(x,y)^2 \Big(\frac{s'_{K_{\min}}(d(x,y))}{s_{K_{\min}}(d(x,y))} \Big)^2, \\
|\operatorname{Div}_x \operatorname{Log}_x(y)|^2 &\le 2 + 2(d-1)^2\, d(x,y)^2 \Big(\frac{s'_{K_{\min}}(d(x,y))}{s_{K_{\min}}(d(x,y))} \Big)^2, \\
\| \operatorname{Grad} \operatorname{Div}_x \operatorname{Log}_x(y) \| &\le d\, \frac{s'_{K_{\min}}(d(x,y))}{s_{K_{\min}}(d(x,y))} \left(d(x,y) \frac{s'_{K_{\min}}(d(x,y))}{s_{K_{\min}}(d(x,y))} - 1 \right) \\
&\quad + d(-K_{\min})\, d(x,y) + \frac{2dL_R}{d(x,y)} s_{K_{\min}}(d(x,y))^2,
\end{aligned}
$$

where on $\mathrm{SPD}\,(n)$, the last inequality can be bounded as

$$
\| \operatorname{Grad} \operatorname{Div}_x \operatorname{Log}_x(y) \| \le \frac{\sqrt{2}d}{2} \big(2\big(1 + d(x,y)\, \frac{s'_{K_{\min}}(d(x,y))}{s_{K_{\min}}(d(x,y))} \big) \big)^{\frac{3}{2}}.
$$

Now we prove the inequalities on derivatives of $\log p_t$. We remark that due to Lemma 38, comparing point-wise upper bounds for derivative of $\log J_t$ and that of $\log p_0$, since the $d(x_0,z)$ term doesn't

increase the order of expectation compared with $d(x_0, x_1)$ term, we can treat $d(x_0, z)$ as $d(x_0, x_1)$ and see that derivatives of $\log J_t$ dominate over the same derivative of $\log p_0$. Thus in computing derivatives of $\log p_t = \log p_0(\Psi_{t,x_1}(x)) + \log J_t(x|x_1) + \text{const}$, it suffices to consider derivatives of $\log J_t(x|x_1)$ terms.

The first four inequalities are straightforward by applying the Cauchy-Schwarz inequality, and note that they are dominated by the $\log J_t$ terms. Following Lemma 27

$$\begin{aligned}\|\nabla_x^2 \log p_t(x_1 \mid x)\|_{\mathrm{op}}^2 \lesssim & \|\nabla^2 \log J_t(x \mid X_1)\|_{\mathrm{op}}^2 + \mathbb{E}\big[\|\nabla^2 \log J_t(x \mid X_1)\|_{\mathrm{op}}^2 \mid X_t = x\big] \\ & + \mathbb{E}\big[\|\operatorname{Grad}_x \log J_t(x \mid X_1)\|^2 \mid X_t = x\big]\mathbb{E}\big[\|\operatorname{Grad} \log p_t(x_1 \mid x)\|^2 \mid X_t = x\big], \\ \|\nabla_x^2 \log p_t(x_1 \mid x)\|_{\mathrm{op}}^4 \lesssim & \|\nabla^2 \log J_t(x \mid X_1)\|_{\mathrm{op}}^4 \\ & + \mathbb{E}\big[\|\operatorname{Grad}_x \log J_t(x \mid X_1)\|^4 \mid X_t = x\big]\mathbb{E}\big[\|\operatorname{Grad} \log p_t(x_1 \mid x)\|^4 \mid X_t = x\big],\end{aligned}$$

and plug in the corresponding expression (using Cauch-Schwarz), we obtain the bounds for the Hessian of $\log p_t$, and following Lemma 28,

$$\begin{aligned}\|\partial_t \operatorname{Grad}_x \log p_t(x_1 \mid x)\| \le & \|\partial_t \operatorname{Grad}_x \log J_t(x \mid x_1)\| + \mathbb{E}\big[\|\partial_t \operatorname{Grad}_x \log J_t(x \mid X_1)\| \mid X_t = x\big] \\ & + \mathbb{E}[\|\operatorname{Grad}_x \log J_t(x \mid X_1)\| \, |\partial_t \log p_t(X_1 \mid x)| \mid X_t = x],\end{aligned}$$

we plug in the corresponding expressions (using Cauch-Schwarz), we obtain the last inequality. □

We summarize the required bound in the following Lemma, which is similar to Lemma 29, 30, 31 and 32.

Lemma 40. *On a Hadamard manifold that is also a symmetric space, with prior distribution being $p_0 \propto \exp(-d d(x_0, z)^2)$, under the conditions in Assumption 1, we have*

$$\begin{aligned}\mathbb{E}[\|v(t,x)\|^2] \lesssim & \mathbb{E}[d(x_0,x_1)^2], \\ \mathbb{E}[\|\nabla v(t,x)\|] \lesssim & \frac{d}{1-t} L_R \mathbb{E}[d(x_0,x_1)^2]^{\frac{1}{2}} \mathbb{E}[s_{K_{\min}}(d(x_0,x_1))^6]^{\frac{1}{2}}, \\ \mathbb{E}[|\frac{d}{dt} v(t,x)|] \lesssim & \frac{d}{1-t} L_R \mathbb{E}[d(x_0,x_1)^2 s_{K_{\min}}(d(x_0,x_1))^6]^{\frac{1}{2}} \mathbb{E}[d(x_0,x_1)^2]^{\frac{1}{2}}, \\ \mathbb{E}[\|\mathit{Grad}_x \mathit{Div}\ v(t,x)\|] \lesssim & \mathbb{E}[d(x_0,x_1) s_{K_{\min}}(d(x_0,x_1))^6] \frac{d^2}{(1-t)^2} L_R^3, \\ \mathbb{E}[|\frac{d}{dt} \mathit{Div}\ v(t,x)|] \lesssim & \mathbb{E}[d(x_0,x_1)^2]^{\frac{1}{2}} \frac{d^2}{(1-t)^2} L_R^3 \mathbb{E}[d(x_1,x_0)^2 s_{K_{\min}}(d(x_0,x_1))^{12}]^{\frac{1}{2}}.\end{aligned}$$

Furthermore,

$$\begin{aligned}\mathbb{E}[\|\nabla v(t,x)\|^2] \lesssim & \frac{d^2}{(1-t)^2} L_R^2 \mathbb{E}[d(x_0,x_1)^4]^{\frac{1}{2}} \mathbb{E}[s_{K_{\min}}(d(x_0,x_1))^{12}]^{\frac{1}{2}}, \\ \mathbb{E}[\|\nabla v(t,x)\|_{\mathrm{HS}}^2] \lesssim & \frac{d^2}{(1-t)^2} L_R^2 \mathbb{E}[d(x_0,x_1)^4]^{\frac{1}{2}} \mathbb{E}[s_{K_{\min}}(d(x_0,x_1))^{12}]^{\frac{1}{2}}, \\ \mathbb{E}[|\frac{d}{dt} v(t,x)|^2] \lesssim & \frac{d^2}{(1-t)^2} L_R^2 \mathbb{E}[d(x_0,x_1)^4 s_{K_{\min}}(d(x_0,x_1))^{12}]^{\frac{1}{2}} \mathbb{E}[d(x_0,x_1)^4]^{\frac{1}{2}}, \\ \mathbb{E}[\|\mathit{Grad}_x \mathit{Div}\ v(t,x)\|^2] \lesssim & \mathbb{E}[d(x_0,x_1)^2 s_{K_{\min}}(d(x_0,x_1))^{12}] \frac{d^4}{(1-t)^4} L_R^6.\end{aligned}$$

Proof. Throughout the proof, we will use the bounds in Lemma 39.We first control the vector field regularity. This is similar to Lemma 29 and Lemma 30. We have

$$\begin{aligned}\mathbb{E}[\|\nabla v(t,x)\|] \le & \frac{1}{1-t} \mathbb{E}\big[\|\operatorname{Log}_x(X_1)\| \, \|\operatorname{Grad}_x \log p_t(X_1 \mid x)\| \mid X_t = x\big] \\ & + \frac{1}{1-t} \mathbb{E}[\|\nabla \operatorname{Log}_x(x_1)\| \mid X_t = x] \\ \lesssim & \frac{d}{1-t} L_R \mathbb{E}[d(x_0,x_1)^2]^{\frac{1}{2}} \mathbb{E}[s_{K_{\min}}(d(x_0,x_1))^6]^{\frac{1}{2}},\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[|\partial_t v(t,x)|] =& \frac{1}{(1-t)^2}\mathbb{E}[\mathrm{Log}_x(x_1) \mid X_t = x] + \frac{1}{1-t}\mathbb{E}[\mathrm{Log}_x(x_1)\, \partial_t \log p_t(x_1 \mid x) \mid X_t = x] \\ \lesssim& \frac{d}{1-t} L_R \mathbb{E}[d(x_0,x_1)^2 s_{K_{\min}}(d(x_0,x_1))^6]^{\frac{1}{2}} \mathbb{E}[d(x_0,x_1)^2]^{\frac{1}{2}}.\end{aligned}$$

Together with $\frac{d}{dt} v(t,x) = \partial_t v(t,x) + \nabla_{v(t,x)} v(t,x)$, we obtain the desired bound.

Now we control the divergence regularity. Similar to Lemma 31 and , we have

$$\| \mathrm{Grad}_x \mathrm{Div}\ v(t,x)\| \le \frac{1}{1-t}\Big(T_1 + T_2 + T_3 + T_4\Big),$$

where

$$\begin{aligned}\mathbb{E}[T_1] =& \mathbb{E}[\| \mathrm{Grad}_x \mathrm{Div}_x \mathrm{Log}_x(X_1)\|] \\ \lesssim& dL_R \mathbb{E}[\frac{1}{d(x,y)} s_{K_{\min}}(d(x,y))^2], \\ \mathbb{E}[T_2] =& \mathbb{E}[|\mathrm{Div}_x \mathrm{Log}_x(X_1)|\, \| \mathrm{Grad}_x \log p_t(X_1 \mid x)\|] \\ \lesssim& \mathbb{E}[|\mathrm{Div}_x \mathrm{Log}_x(X_1)|^2]^{\frac{1}{2}} \mathbb{E}[\| \mathrm{Grad}_x \log p_t(X_1 \mid x)\|^2]^{\frac{1}{2}} \\ \lesssim& \frac{d^2}{1-t} L_R \mathbb{E}[d(x_0,x_1)^2]^{\frac{1}{2}} \mathbb{E}[s_{K_{\min}}(d(x_0,x_1))^6]^{\frac{1}{2}}, \\ \mathbb{E}[T_3] =& \mathbb{E}[\| \mathrm{Grad}_x \langle \mathrm{Grad}_x \log p_t(X_1 \mid x), \mathrm{Log}_x(X_1)\rangle\|] \\ \le& \mathbb{E}[\|\nabla \mathrm{Log}_x(X_1)\|\, \| \mathrm{Grad}_x \log p_t(X_1 \mid x)\|] \\ & + \mathbb{E}[\|\nabla^2 \log p_t(X_1 \mid x)\|\, \| \mathrm{Log}_x(X_1)\|] \\ \lesssim& \Big(\mathbb{E}[d(x,x_1)^2] \frac{d^2}{(1-t)^2} L_R^2 \mathbb{E}[s_{K_{\min}}(d(x_0,x_1))^6]\Big)^{\frac{1}{2}} \\ & + \Big(\frac{d^2}{(1-t)^4} L_R^6 \mathbb{E}[s_{K_{\min}}(d(x_0,x_1))^{12}] \mathbb{E}[d(x,x_1)^2]\Big)^{\frac{1}{2}} \\ \lesssim& \frac{d}{1-t} L_R^3 \mathbb{E}[s_{K_{\min}}(d(x_0,x_1))^{12}]^{\frac{1}{2}} \mathbb{E}[d(x_0,x_1)^2]^{\frac{1}{2}}, \\ \mathbb{E}[T_4] =& \mathbb{E}[|\langle \mathrm{Grad}_x \log p_t(X_1 \mid x), \mathrm{Log}_x(X_1)\rangle|\, \| \mathrm{Grad}_x \log p_t(X_1 \mid x)\|] \\ \le& \mathbb{E}[\| \mathrm{Log}_x(X_1)\|\, \| \mathrm{Grad}_x \log p_t(X_1 \mid x)\|^2] \\ \lesssim& \mathbb{E}[d(x,x_1) s_{K_{\min}}(d(x_0,x_1))^6] \frac{d^2}{(1-t)^2} L_R^2 \\ \lesssim& \mathbb{E}[d(x_0,x_1) s_{K_{\min}}(d(x_0,x_1))^6] \frac{d^2}{1-t} L_R^2.\end{aligned}$$

Similar to Lemma 32,

$$\begin{aligned}\frac{d}{dt} \mathrm{Div}\ v(t,x) =& \frac{1}{(1-t)^2}\mathbb{E}\big[\mathrm{Div}_x \mathrm{Log}_x(X_1) + \langle \mathrm{Grad}_x \log p_t(X_1 \mid x), \mathrm{Log}_x(X_1)\rangle \mid X_t = x\big] \\ & + \frac{1}{1-t}\mathbb{E}[\mathrm{Div}_x \mathrm{Log}_x(X_1)\, \partial_t \log p_t(X_1 \mid x) \mid X_t = x] \\ & + \frac{1}{1-t}\mathbb{E}[\partial_t \langle \mathrm{Grad}_x \log p_t(X_1 \mid x), \mathrm{Log}_x(X_1)\rangle \mid X_t = x] \\ & + \frac{1}{1-t}\mathbb{E}[\langle \mathrm{Grad}_x \log p_t(X_1 \mid x), \mathrm{Log}_x(X_1)\rangle\, \partial_t \log p_t(X_1 \mid x) \mid X_t = x].\end{aligned}$$

We bound

$$\begin{aligned}\mathbb{E}[T_1] =& \frac{1}{(1-t)^2}\mathbb{E}\big[\mathrm{Div}_x \mathrm{Log}_x(X_1) + \langle \mathrm{Grad}_x \log p_t(X_1 \mid x), \mathrm{Log}_x(X_1)\rangle\big] \\ \lesssim& \frac{d}{(1-t)^2}\mathbb{E}[d(x,x_1)] + \frac{1}{(1-t)^2}\mathbb{E}[d(x,x_1)^2]^{\frac{1}{2}} \mathbb{E}[\frac{d^2}{(1-t)^2} L_R^2 s_{K_{\min}}(d(x_0,x_1))^6]^{\frac{1}{2}}\end{aligned}$$

$$\lesssim \frac{d}{1-t}\mathbb{E}[d(x_0,x_1)] + \frac{d}{(1-t)^2}L_R\mathbb{E}[d(x_0,x_1)^2]^{\frac{1}{2}}\mathbb{E}[s_{K_{\min}}(d(x_0,x_1))^6]]^{\frac{1}{2}},$$

$$\begin{aligned}
\mathbb{E}[T_2] =& \frac{1}{1-t}\mathbb{E}[\text{Div}_x \operatorname{Log}_x(X_1)\,\partial_t \log p_t(X_1 \mid x)] \\
\lesssim& \frac{1}{1-t} d\mathbb{E}[d(x,x_1)^2]^{\frac{1}{2}} \frac{d}{1-t} L_R\mathbb{E}[d(x_0,x_1)^2 s_{K_{\min}}(d(x_0,x_1))^6]^{\frac{1}{2}} \\
\lesssim& \frac{d^2}{1-t} L_R\mathbb{E}[d(x_0,x_1)^2]^{\frac{1}{2}}\mathbb{E}[d(x_0,x_1)^2 s_{K_{\min}}(d(x_0,x_1))^6]^{\frac{1}{2}}, \\
\mathbb{E}[T_3] =& \frac{1}{1-t}\mathbb{E}[\partial_t\langle \text{Grad}_x \log p_t(X_1 \mid x), \operatorname{Log}_x(X_1)\rangle] \\
\lesssim& \mathbb{E}[d(x_0,x_1)^2]^{\frac{1}{2}} \frac{d}{(1-t)^2} L_R^3\mathbb{E}[d(x_1,x_0)^2 s_{K_{\min}}(d(x_0,x_1))^{12}]^{\frac{1}{2}}, \\
\mathbb{E}[T_4] =& \frac{1}{1-t}\mathbb{E}[\langle \text{Grad}_x \log p_t(X_1 \mid x), \operatorname{Log}_x(X_1)\rangle\, \partial_t \log p_t(X_1 \mid x)] \\
\lesssim& \frac{1}{1-t}\Big(\frac{d^3}{(1-t)^3} L_R^3\mathbb{E}[s_{K_{\min}}(d(x_0,x_1))^9]\Big)^{\frac{1}{3}}\mathbb{E}[d(x,x_1)^3]^{\frac{1}{3}}\Big(\frac{d^3}{(1-t)^3} L_R^3\mathbb{E}[d(x_0,x_1)^3 s_{K_{\min}}(d(x_0,x_1))^9]\Big)^{\frac{1}{3}} \\
\lesssim& \frac{d^2}{(1-t)^2} L_R^2\mathbb{E}[s_{K_{\min}}(d(x_0,x_1))^9]^{\frac{1}{3}}\mathbb{E}[d(x_0,x_1)^3]^{\frac{1}{3}}\mathbb{E}[d(x_0,x_1)^3 s_{K_{\min}}(d(x_0,x_1))^9]^{\frac{1}{3}}.
\end{aligned}$$

We also have

$$\mathbb{E}[|\langle \text{Grad}_x \text{ Div } v(t,x), v(t,x)\rangle|] \lesssim \frac{d^2 L_R^3}{(1-t)^2}\mathbb{E}[d(x_0,x_1)^2 s_{K_{\min}}(d(x_0,x_1))^6].$$

Using $\frac{d}{dt}\text{ Div } v(t,x) = \partial_t \text{ Div } v(t,x) + \langle \text{Grad}_x \text{ Div } v(t,x), v(t,x)\rangle$ we get the desired bound.

Higher order bounds and $\mathbb{E}\|\nabla v(t,X_t)\|_{\text{HS}}^2$ follow exactly the same procedure.

□

E.3 Auxiliary Bounds on Derivatives

Recall that we are using the following notation

$$\Psi_{t,x_1}(x) = \operatorname{Exp}_{x_1}\Big(\tfrac{1}{1-t}\operatorname{Log}_{x_1}(x)\Big).$$

Lemma 41 (Controlling derivatives of $\log p_0$)**.** *Let M be a Hadamard manifold. Fix $\beta > 0$, $m = 2$, and $z \in M$, and define*

$$p_0(x) \propto \exp\big(-\beta\, d(x,z)^m\big).$$

Fix $t \in [0,1)$, $x_1 \in M$, and $x \in M$. Define $r := d(\Psi_{t,x_1}(x), z)$, so that r represent the radial distance between x_0 and z. Then the following bounds hold:

$$\begin{aligned}
\Big\|\mathit{Grad}_x \log p_0 \Psi_{t,x_1}(x)\Big\| \leq& 2\beta r\Big\|(d\mathit{Exp}_{x_1})_{\frac{1}{1-t}\mathit{Log}_{x_1}(x)}\Big\|_{\text{op}}\ \frac{1}{1-t}\ \big\|(d\mathit{Log}_{x_1})_x\big\|_{\text{op}}, \\
|\partial_t \log p_0(\Psi_{t,x_1}(x))| \leq& 2\beta r\Big\|(d\mathit{Exp}_{x_1})_{\frac{1}{1-t}\mathit{Log}_{x_1}(x)}\Big\|_{\text{op}}\ \frac{1}{(1-t)^2}\ \big\|\mathit{Log}_{x_1}(x)\big\|, \\
\|\nabla_x^2 \log p_0(\Psi_{t,x_1}(x))\|_{\text{op}} \leq& \Big(\Big\|(d\mathit{Exp}_{x_1})_{\frac{1}{1-t}\mathit{Log}_{x_1}(x)}\Big\|_{\text{op}}\ \frac{1}{1-t}\ \big\|(d\mathit{Log}_{x_1})_x\big\|_{\text{op}}\Big)^2 \\
&\Big(2\beta + 2\beta r\|\nabla^2 r\|_{\text{op}}\Big|_{\Psi_{t,x_1}(x)}\Big) \\
&+ \Big\|\nabla(d\mathit{Exp}_{x_1})_{\frac{1}{1-t}\mathit{Log}_{x_1}(x)}\Big\|\Big(\frac{1}{1-t}\big\|(d\mathit{Log}_{x_1})_x\big\|_{\text{op}}\Big)^2 2\beta r \\
&+ \Big\|(d\mathit{Exp}_{x_1})_{\frac{1}{1-t}\mathit{Log}_{x_1}(x)}\Big\|_{\text{op}}\ \frac{1}{1-t}\ \big\|\nabla(d\mathit{Log}_{x_1})_x\big\|\, 2\beta r,
\end{aligned}$$

$$\begin{aligned}\Big\|\partial_t \mathit{Grad}_x \log p_0(\Psi_{t,x_1}(x))\Big\| \le& \Big(\Big\|(d\,\mathit{Exp}_{x_1})_{\frac{1}{1-t}\mathit{Log}_{x_1}(x)}\Big\|_{\mathrm{op}}\ \frac{1}{1-t}\ \big\|(d\,\mathit{Log}_{x_1})_x\big\|_{\mathrm{op}}\Big)\\ &\times\Big(2\beta+2\beta r\|\nabla^2 r\|_{\mathrm{op}}\Big|_{\Psi_{t,x_1}(x)}\Big)\\ &\times\Big\|(d\,\mathit{Exp}_{x_1})_{\frac{1}{1-t}\mathit{Log}_{x_1}(x)}\Big\|_{\mathrm{op}}\ \frac{1}{(1-t)^2}\ \big\|\mathit{Log}_{x_1}(x)\big\|\\ &+\Bigg[\frac{1}{(1-t)^2}\Big\|(d\,\mathit{Exp}_{x_1})_{\frac{1}{1-t}\mathit{Log}_{x_1}(x)}\Big\|_{\mathrm{op}}\ \big\|(d\,\mathit{Log}_{x_1})_x\big\|_{\mathrm{op}}\\ &+\frac{1}{(1-t)^3}\ \big\|\mathit{Log}_{x_1}(x)\big\|\ \Big\|\nabla(d\,\mathit{Exp}_{x_1})_{\frac{1}{1-t}\mathit{Log}_{x_1}(x)}\Big\|\ \big\|(d\,\mathit{Log}_{x_1})_x\big\|_{\mathrm{op}}\Bigg]2\beta r.\end{aligned}$$

where

$$\|\nabla^2 r\|_{\mathrm{op}}\Big|_{\Psi_{t,x_1}(x)} \le \frac{s'_{K_{\min}}(r)}{s_{K_{\min}}(r)} = \sqrt{-K_{\min}}\coth(r\sqrt{-K_{\min}}).$$

and note that $r\mapsto r\cosh(r)$ has no singularity at $r=0$.

Proof. We start with deriving bounds for $\|\operatorname{Grad}\log p_0(\Psi_{t,x_1}(x))\|$ and $\|\nabla^2\log p_0(\Psi_{t,x_1}(x))\|_{\mathrm{op}}$. Since $\log p_0(y) = -\beta\, d(y,z)^m + \text{const}$, and $\|\operatorname{Grad} d(\,\cdot\,,z)\| = 1$, we have

$$\|\operatorname{Grad}\log p_0(\Psi_{t,x_1}(x))\| = 2\beta r.$$

Moreover, notice that

$$\nabla^2(d(y,z)^m) = 2d(y,z)\nabla^2 d(y,z) + 2\operatorname{Grad} d(y,z)\otimes\operatorname{Grad} d(y,z).$$

The above identity, together with $\|\operatorname{Grad} d(y,z)\otimes\operatorname{Grad} d(y,z)\|_{\mathrm{op}} = 1$, yields

$$\|\nabla^2\log p_0(\Psi_{t,x_1}(x))\|_{\mathrm{op}} \le 2\beta + 2\beta\, r\,\|\nabla^2 r\|_{\mathrm{op}}\Big|_{\Psi_{t,x_1}(x)}.$$

We next derive the first (spatial) derivative of Ψ_{t,x_1}. Using the chain rule,

$$(d\Psi_{t,x_1})_x[u] = (d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\Big[\tfrac{1}{1-t}\,(d\operatorname{Log}_{x_1})_x[u]\Big],\qquad \forall\, u\in T_xM.$$

Hence

$$\|(d\Psi_{t,x_1})_x\|_{\mathrm{op}} \le \Big\|(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\Big\|_{\mathrm{op}}\ \tfrac{1}{1-t}\ \big\|(d\operatorname{Log}_{x_1})_x\big\|_{\mathrm{op}}. \tag{25}$$

For any $u\in T_xM$, the chain rule gives the identity

$$\big\langle \operatorname{Grad}_x\log p_0(\Psi_{t,x_1}(x)),\, u\big\rangle = \big\langle \operatorname{Grad}\log p_0(\Psi_{t,x_1}(x)),\,(d\Psi_{t,x_1})_x[u]\big\rangle.$$

Therefore, by Cauchy-Schwarz with supremum over $\|u\|=1$, yields

$$\begin{aligned}\big|\langle\operatorname{Grad}_x\log p_0(\Psi_{t,x_1}(x)),u\rangle\big| &\le \|\operatorname{Grad}\log p_0(\Psi_{t,x_1}(x))\|\,\|(d\Psi_{t,x_1})_x[u]\|\\ &\le \|\operatorname{Grad}\log p_0(\Psi_{t,x_1}(x))\|\,\|(d\Psi_{t,x_1})_x\|_{\mathrm{op}}.\end{aligned}$$

Substituting $\|\operatorname{Grad}\log p_0(\Psi_{t,x_1}(x))\| = 2\beta r$ and using (25), we obtain

$$\begin{aligned}&\Big\|\operatorname{Grad}_x\log p_0\Big(\operatorname{Exp}_{x_1}\big(\tfrac{1}{1-t}\operatorname{Log}_{x_1}(x)\big)\Big)\Big\|\\ \le\ & 2\beta r\,\Big\|(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\Big\|_{\mathrm{op}}\ \frac{1}{1-t}\ \big\|(d\operatorname{Log}_{x_1})_x\big\|_{\mathrm{op}}.\end{aligned}$$

Now we work on the first time derivative. Recall $\Psi_{t,x_1}(x) = \operatorname{Exp}_{x_1}\Big(\frac{1}{1-t}\operatorname{Log}_{x_1}(x)\Big)$. By chain rule,

$$\partial_t\log p_0(\Psi_{t,x_1}(x)) = \big\langle\operatorname{Grad}\log p_0(\Psi_{t,x_1}(x)),\,\partial_t\Psi_{t,x_1}(x)\big\rangle,$$

where $\partial_t \Psi_{t,x_1}(x) = (d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\Big[\frac{1}{(1-t)^2}\operatorname{Log}_{x_1}(x)\Big]$. By Cauchy-Schwarz,

$$\begin{aligned}
|\partial_t \log p_0(\Psi_{t,x_1}(x))| &\le \|\operatorname{Grad}\, \log p_0(\Psi_{t,x_1}(x))\|\, \|\partial_t \Psi_{t,x_1}(x)\| \\
&\le \|\operatorname{Grad}\, \log p_0(\Psi_{t,x_1}(x))\| \, \Big\|(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\Big\|_{\mathrm{op}}\, \tfrac{1}{(1-t)^2}\, \|\operatorname{Log}_{x_1}(x)\| \\
&\le 2\beta r \Big\|(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\Big\|_{\mathrm{op}}\, \frac{1}{(1-t)^2}\, \big\|\operatorname{Log}_{x_1}(x)\big\|,
\end{aligned}$$

where recall $\|\operatorname{Grad}\, \log p_0(\Psi_{t,x_1}(x))\| = 2\beta r$.

Now we compute the Hessian. For any $u, w \in T_x M$, using chain rule, we obtain

$$\begin{aligned}
\nabla_x^2 \log p_0(\Psi_{t,x_1}(x))[u,w] =& \nabla^2 \log p_0(\Psi_{t,x_1}(x))\big[(\Psi_{t,x_1})_x[u], (\Psi_{t,x_1})_x[w]\big] \\
&+ \big\langle \operatorname{Grad}\, \log p_0(\Psi_{t,x_1}(x)),\, \nabla_u\big((\Psi_{t,x_1})_x[w]\big)\big\rangle.
\end{aligned}$$

By Cauchy-Schwarz and taking the supremum over $\|u\| = \|w\| = 1$, we obtain

$$\begin{aligned}
\|\nabla_x^2 \log p_0(\Psi_{t,x_1}(x))\|_{\mathrm{op}} &= \sup_{\|u\|=\|w\|=1} \big|\nabla_x^2 \log p_0(\Psi_{t,x_1}(x))[u,w]\big| \\
\le& \sup_{\|u\|=\|w\|=1} \|\nabla^2 \log p_0(\Psi_{t,x_1}(x))\|_{\mathrm{op}}\, \|(\Psi_{t,x_1})_x[u]\|\, \|(\Psi_{t,x_1})_x[w]\| \\
&+ \|\operatorname{Grad}\, \log p_0(\Psi_{t,x_1}(x))\|\, \big\|\nabla_u\big((\Psi_{t,x_1})_x[w]\big)\big\| \\
\le& \|\nabla^2 \log p_0(\Psi_{t,x_1}(x))\|_{\mathrm{op}}\, \|(\Psi_{t,x_1})_x\|_{\mathrm{op}}^2 + \|\operatorname{Grad}\, \log p_0(\Psi_{t,x_1}(x))\|\, \|\nabla(\Psi_{t,x_1})_x\|.
\end{aligned}$$

It remains to bound $\|\nabla(\Psi_{t,x_1})_x\|$. Recall that

$$(\Psi_{t,x_1})_x[w] = (d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\Big[\tfrac{1}{1-t}(d\operatorname{Log}_{x_1})_x[w]\Big].$$

Recall the Leibniz rule with $v = \frac{1}{1-t}\operatorname{Log}_{x_1}(x)$ being a x-dependent tangent vector at $T_{x_1}M$.

$$\begin{aligned}
\nabla_u\big((\Psi_{t,x_1})_x[w]\big) &= \nabla_u\Big((d\operatorname{Exp}_{x_1})_v\Big[\tfrac{1}{1-t}(d\operatorname{Log}_{x_1})_x[w]\Big]\Big) \\
&= \Big(\nabla_{\nabla_u v}(d\operatorname{Exp}_{x_1})\Big)_v\Big[\tfrac{1}{1-t}(d\operatorname{Log}_{x_1})_x[w]\Big] \;+\; (d\operatorname{Exp}_{x_1})_v\Big[\tfrac{1}{1-t}\,\nabla_u\big((d\operatorname{Log}_{x_1})_x[w]\big)\Big].
\end{aligned}$$

where note that v depends on x, so using chain rule,

$$\begin{aligned}
\nabla_u((d\operatorname{Exp}_{x_1})_v) &= \nabla_u((d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}) \\
&= \big(\nabla(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\big)(\nabla_u \tfrac{1}{1-t}\operatorname{Log}_{x_1}(x)) \\
&= \Big(\nabla_{\nabla_u \frac{1}{1-t}\operatorname{Log}_{x_1}(x)}(d\operatorname{Exp}_{x_1})\Big)_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}.
\end{aligned}$$

Notice that for $v = \frac{1}{1-t}\operatorname{Log}_{x_1}(x)$, its first derivative is a directional derivative. We have

$$\nabla_u v = \tfrac{1}{1-t}(d\operatorname{Log}_{x_1})_x[u], \qquad \text{and} \qquad \nabla_u\big((d\operatorname{Log}_{x_1})_x[w]\big) = \nabla(d\operatorname{Log}_{x_1})_x[u,w].$$

Taking norms and applying the definitions of the operator norms gives

$$\begin{aligned}
\big\|\nabla_u\big((\Psi_{t,x_1})_x[w]\big)\big\| &\le \Big\|\nabla(d\operatorname{Exp}_{x_1})_v\Big\|\, \|\nabla_u v\|\, \Big\|\tfrac{1}{1-t}(d\operatorname{Log}_{x_1})_x[w]\Big\| \\
&\quad + \Big\|(d\operatorname{Exp}_{x_1})_v\Big\|_{\mathrm{op}}\, \tfrac{1}{1-t}\, \big\|\nabla(d\operatorname{Log}_{x_1})_x[u,w]\big\| \\
&= \Big\|\nabla(d\operatorname{Exp}_{x_1})_v\Big\|\, \Big\|\tfrac{1}{1-t}(d\operatorname{Log}_{x_1})_x[u]\Big\|\, \Big\|\tfrac{1}{1-t}(d\operatorname{Log}_{x_1})_x[w]\Big\| \\
&\quad + \Big\|(d\operatorname{Exp}_{x_1})_v\Big\|_{\mathrm{op}}\, \tfrac{1}{1-t}\, \big\|\nabla(d\operatorname{Log}_{x_1})_x[u,w]\big\|.
\end{aligned}$$

Taking supremum over $\|u\| = \|w\| = 1$, we obtain

$$\big\|\nabla\big((d\Psi_{t,x_1})_x\big)\big\| \le \Big\|\nabla(d\operatorname{Exp}_{x_1})_v\Big\|\, \Big(\tfrac{1}{1-t}\big\|(d\operatorname{Log}_{x_1})_x\big\|_{\mathrm{op}}\Big)^2 + \Big\|(d\operatorname{Exp}_{x_1})_v\Big\|_{\mathrm{op}}\, \tfrac{1}{1-t}\, \big\|\nabla(d\operatorname{Log}_{x_1})_x\big\|.$$

Finally using the fact that

$$\begin{aligned}\|\operatorname{Grad}\,\log p_0(\Psi_{t,x_1}(x))\| &= 2\beta r,\\ \|\nabla^2 \log p_0(\Psi_{t,x_1}(x))\|_{\mathrm{op}} &\le 2\beta + 2\beta r\|\nabla^2 r\|_{\mathrm{op}}\big|_{\Psi_{t,x_1}(x)},\end{aligned}$$

and using (25), we have

$$\begin{aligned}&\|\nabla_x^2 \log p_0(\Psi_{t,x_1}(x))\|_{\mathrm{op}}\\ \le&\|\nabla^2 \log p_0(\Psi_{t,x_1}(x))\|_{\mathrm{op}}\ \|(\Psi_{t,x_1})_x\|_{\mathrm{op}}^2 + \|\operatorname{Grad}\,\log p_0(\Psi_{t,x_1}(x))\|\,\|\nabla(\Psi_{t,x_1})_x\|\\ \le&\left(\left\|(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\right\|_{\mathrm{op}}\ \frac{1}{1-t}\ \left\|(d\operatorname{Log}_{x_1})_x\right\|_{\mathrm{op}}\right)^2\left(2\beta + 2\beta r\|\nabla^2 r\|_{\mathrm{op}}\Big|_{\Psi_{t,x_1}(x)}\right)\\ &+\left\|\nabla(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\right\|\left(\frac{1}{1-t}\left\|(d\operatorname{Log}_{x_1})_x\right\|_{\mathrm{op}}\right)^2 2\beta r\\ &+\left\|(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\right\|_{\mathrm{op}}\ \frac{1}{1-t}\left\|\nabla(d\operatorname{Log}_{x_1})_x\right\| 2\beta r.\end{aligned}$$

It remains to consider the mixed derivative term. For any $u \in T_xM$, recall we have

$$\big\langle \operatorname{Grad}_x \log p_0(\Psi_{t,x_1}(x)), u\big\rangle = \big\langle \operatorname{Grad}\,\log p_0(\Psi_{t,x_1}(x)), (\Psi_{t,x_1})_x[u]\big\rangle.$$

Differentiate both sides in t (with x_1, x, u fixed):

$$\begin{aligned}\big\langle \partial_t \operatorname{Grad}_x \log p_0(\Psi_{t,x_1}(x)), u\big\rangle =&\Big\langle \nabla^2 \log p_0(\Psi_{t,x_1}(x))\big[\partial_t\Psi_{t,x_1}(x)\big],\ (\Psi_{t,x_1})_x[u]\Big\rangle\\ &+\big\langle \operatorname{Grad}\,\log p_0(\Psi_{t,x_1}(x)),\ \partial_t\big((\Psi_{t,x_1})_x[u]\big)\big\rangle.\end{aligned}$$

By Cauchy-Schwarz,

$$\begin{aligned}\big|\langle \partial_t \operatorname{Grad}_x \log p_0(\Psi_{t,x_1}(x)), u\rangle\big| \le&\ \|\nabla^2 \log p_0(\Psi_{t,x_1}(x))\|_{\mathrm{op}}\ \|\partial_t\Psi_{t,x_1}(x)\|\,\|(\Psi_{t,x_1})_x\|_{\mathrm{op}}\ \|u\|\\ &+\|\operatorname{Grad}\,\log p_0(\Psi_{t,x_1}(x))\|\,\Big\|\partial_t\big((\Psi_{t,x_1})_x[u]\big)\Big\|.\end{aligned}$$

We already have $\|\partial_t\Psi_{t,x_1}(x)\|$ and $\|(\Psi_{t,x_1})_x\|_{\mathrm{op}}$ from previous steps.

It remains to bound $\|\partial_t((\Psi_{t,x_1})_x[u])\|$. Recall for any $u \in T_xM$,

$$(\Psi_{t,x_1})_x[u] = (d\operatorname{Exp}_{x_1})_v\Big[\tfrac{1}{1-t}(d\operatorname{Log}_{x_1})_x[u]\Big], \qquad v = \tfrac{1}{1-t}\operatorname{Log}_{x_1}(x).$$

Fix x_1, x, u and differentiate with respect to t. Since x is fixed, the tensor $(d\operatorname{Log}_{x_1})_x[u]$ does not depend on t; only the scalar factor $\frac{1}{1-t}$ and the point v depend on t. By the Leibniz rule (product rule for a t-dependent linear map applied to a t-dependent vector),

$$\partial_t\big((\Psi_{t,x_1})_x[u]\big) = \partial_t\Big((d\operatorname{Exp}_{x_1})_v\Big)\Big[\tfrac{1}{1-t}(d\operatorname{Log}_{x_1})_x[u]\Big] + (d\operatorname{Exp}_{x_1})_v\Big[\partial_t\Big(\tfrac{1}{1-t}(d\operatorname{Log}_{x_1})_x[u]\Big)\Big].$$

For the first term, $(d\operatorname{Exp}_{x_1})_v$ depends on t only through $v(t)$, hence the chain rule gives

$$\partial_t\Big((d\operatorname{Exp}_{x_1})_v\Big) = \Big(\nabla(d\operatorname{Exp}_{x_1})_v\Big)(\partial_t v) = \Big(\nabla_{\partial_t v}(d\operatorname{Exp}_{x_1})\Big)_v.$$

For the second term, since $(d\operatorname{Log}_{x_1})_x[u]$ is t-independent,

$$\partial_t\Big(\tfrac{1}{1-t}(d\operatorname{Log}_{x_1})_x[u]\Big) = \tfrac{1}{(1-t)^2}(d\operatorname{Log}_{x_1})_x[u].$$

Combining these identities yields the formula

$$\partial_t\big((\Psi_{t,x_1})_x[u]\big) = \Big(\nabla_{\partial_t v}(d\operatorname{Exp}_{x_1})\Big)_v\Big[\tfrac{1}{1-t}(d\operatorname{Log}_{x_1})_x[u]\Big] + (d\operatorname{Exp}_{x_1})_v\Big[\tfrac{1}{(1-t)^2}(d\operatorname{Log}_{x_1})_x[u]\Big].$$

Noting $\partial_t v = \frac{1}{(1-t)^2}\operatorname{Log}_{x_1}(x)$, we take norms and obtain (for $\|u\| = 1$)

$$\Big\|\partial_t\big((\Psi_{t,x_1})_x[u]\big)\Big\|$$

$$
\begin{aligned}
&\leq \Big\| \nabla (d\operatorname{Exp}_{x_1})_v \Big\| \ \|\partial_t v\| \ \Big\| \tfrac{1}{1-t} (d\operatorname{Log}_{x_1})_x[u] \Big\| \\
&\qquad + \Big\| (d\operatorname{Exp}_{x_1})_v \Big\|_{\mathrm{op}} \ \tfrac{1}{(1-t)^2} \ \big\| (d\operatorname{Log}_{x_1})_x \big\|_{\mathrm{op}} \ \|u\| \\
&= \Big\| \nabla (d\operatorname{Exp}_{x_1})_v \Big\| \ \tfrac{1}{(1-t)^2} \| \operatorname{Log}_{x_1}(x)\| \ \tfrac{1}{1-t} \big\| (d\operatorname{Log}_{x_1})_x \big\|_{\mathrm{op}} \\
&\qquad + \Big\| (d\operatorname{Exp}_{x_1})_v \Big\|_{\mathrm{op}} \ \tfrac{1}{(1-t)^2} \ \big\| (d\operatorname{Log}_{x_1})_x \big\|_{\mathrm{op}} \\
&= \tfrac{1}{(1-t)^3} \| \operatorname{Log}_{x_1}(x)\| \Big\| \nabla (d\operatorname{Exp}_{x_1})_v \Big\| \ \big\| (d\operatorname{Log}_{x_1})_x \big\|_{\mathrm{op}} \\
&\qquad + \tfrac{1}{(1-t)^2} \Big\| (d\operatorname{Exp}_{x_1})_v \Big\|_{\mathrm{op}} \ \big\| (d\operatorname{Log}_{x_1})_x \big\|_{\mathrm{op}} .
\end{aligned}
$$

Substituting the expressions for $\| \operatorname{Grad} \log p_0(\Psi_{t,x_1}(x))\|$ and $\|\nabla^2 \log p_0(\Psi_{t,x_1}(x))\|_{\mathrm{op}}$, we get

$$
\begin{aligned}
&\Big\| \partial_t \operatorname{Grad}_x \log p_0 \Big(\operatorname{Exp}_{x_1} \big(\tfrac{1}{1-t} \operatorname{Log}_{x_1}(x) \big) \Big) \Big\| \\
&\leq \Big(\Big\| (d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} \Big\|_{\mathrm{op}} \ \frac{1}{1-t} \ \big\| (d\operatorname{Log}_{x_1})_x \big\|_{\mathrm{op}} \Big) \\
&\quad \times \Big(2\beta + 2\beta r \|\nabla^2 r\|_{\mathrm{op}} \Big|_{\Psi_{t,x_1}(x)} \Big) \\
&\quad \times \Big\| (d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} \Big\|_{\mathrm{op}} \ \frac{1}{(1-t)^2} \ \| \operatorname{Log}_{x_1}(x)\| \\
&\quad + \Bigg[\frac{1}{(1-t)^2} \Big\| (d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} \Big\|_{\mathrm{op}} \ \big\| (d\operatorname{Log}_{x_1})_x \big\|_{\mathrm{op}} \\
&\quad + \frac{1}{(1-t)^3} \| \operatorname{Log}_{x_1}(x)\| \Big\| \nabla (d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} \Big\| \ \big\| (d\operatorname{Log}_{x_1})_x \big\|_{\mathrm{op}} \Bigg] 2\beta r.
\end{aligned}
$$

Finally, we remark that when $K_{\min} < 0$, the Hessian comparison theorem gives

$$
\|\nabla^2 r\|_{\mathrm{op}} \Big|_{\Psi_{t,x_1}(x)} \leq \frac{s'_{K_{\min}}(r)}{s_{K_{\min}}(r)} = \sqrt{-K_{\min}} \coth(r\sqrt{-K_{\min}}).
$$

□

Lemma 42 (Controlling derivatives of J_t-terms)**.** *Assume M is a Hadamard manifold. Fix $x_1 \in M$, $x \in M$, and $t \in (0,1)$. Recall*

$$
\begin{aligned}
J_t(x \mid x_1) &= (1-t)^{-d} \, \frac{\big| \det (d\mathit{Exp}_{x_1})_{\frac{1}{1-t}\mathit{Log}_{x_1}(x)} \big|}{\big| \det (d\mathit{Exp}_{x_1})_{\mathit{Log}_{x_1}(x)} \big|}, \\
\log J_t(x \mid x_1) &= -d \log(1-t) + \log \det (d\mathit{Exp}_{x_1})_{\frac{1}{1-t}\mathit{Log}_{x_1}(x)} - \log \det (d\mathit{Exp}_{x_1})_{\mathit{Log}_{x_1}(x)}.
\end{aligned}
$$

Then the following pointwise bounds hold.

$$
\begin{aligned}
\| \mathit{Grad}_x \log J_t(x \mid x_1)\| \leq & d \, \|(d\mathit{Log}_{x_1})_x\| \left(\frac{1}{1-t} \, \|\nabla (d\mathit{Exp}_{x_1})_{\frac{1}{1-t}\mathit{Log}_{x_1}(x)}\| + \|\nabla (d\mathit{Exp}_{x_1})_{\mathit{Log}_{x_1}(x)}\| \right), \\
\|\nabla_x^2 \log J_t(x \mid x_1)\|_{\mathrm{op}} \leq & d \, \|(d\mathit{Log}_{x_1})_x\|^2 \Bigg[\frac{1}{(1-t)^2} \Big(\|\nabla (d\mathit{Exp}_{x_1})_{\frac{1}{1-t}\mathit{Log}_{x_1}(x)}\|^2 \\
& + \|\nabla^2 (d\mathit{Exp}_{x_1})_{\frac{1}{1-t}\mathit{Log}_{x_1}(x)}\| \Big) \\
& + \Big(\|\nabla (d\mathit{Exp}_{x_1})_{\mathit{Log}_{x_1}(x)}\|^2 + \|\nabla^2 (d\mathit{Exp}_{x_1})_{\mathit{Log}_{x_1}(x)}\| \Big) \Bigg] \\
& + d \, \|\nabla (d\mathit{Log}_{x_1})_x\| \left(\frac{1}{1-t} \, \|\nabla (d\mathit{Exp}_{x_1})_{\frac{1}{1-t}\mathit{Log}_{x_1}(x)}\| + \|\nabla (d\mathit{Exp}_{x_1})_{\mathit{Log}_{x_1}(x)}\| \right),
\end{aligned}
$$

$$|\partial_t \log J_t(x \mid x_1)| \leq \frac{d}{1-t} + \frac{d}{(1-t)^2} \left\| \mathit{Log}_{x_1}(x) \right\| \left\| \nabla (d\,\mathit{Exp}_{x_1})_{\frac{1}{1-t}\mathit{Log}_{x_1}(x)} \right\|,$$

$$\begin{aligned}\left\| \partial_t \mathit{Grad}_x \log J_t(x \mid x_1) \right\| \leq & \|(d\,\mathit{Log}_{x_1})_x\| \Bigg[\frac{d}{(1-t)^2} \|\nabla (d\,\mathit{Exp}_{x_1})_{\frac{1}{1-t}\mathit{Log}_{x_1}(x)}\| \\ & + \frac{d}{(1-t)^3} \left\| \mathit{Log}_{x_1}(x) \right\| \\ & \left(\|\nabla (d\,\mathit{Exp}_{x_1})_{\frac{1}{1-t}\mathit{Log}_{x_1}(x)}\|^2 + \|\nabla^2 (d\,\mathit{Exp}_{x_1})_{\frac{1}{1-t}\mathit{Log}_{x_1}(x)}\| \right) \Bigg].\end{aligned}$$

Proof. We repeatedly use Lemma 60 with $A(\cdot)$ chosen to be $\xi \mapsto (d\operatorname{Exp}_{x_1})_\xi$, and the chain rule through Log_{x_1}.

We first estimate the gradient. From

$$\log J_t(x \mid x_1) = -d \log(1-t) + \log\det (d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} - \log\det (d\operatorname{Exp}_{x_1})_{\operatorname{Log}_{x_1}(x)},$$

the constant $-d\log(1-t)$ has zero x-gradient, hence

$$\operatorname{Grad}_x \log J_t(x \mid x_1) = \operatorname{Grad}_x \log\det (d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} - \operatorname{Grad}_x \log\det (d\operatorname{Exp}_{x_1})_{\operatorname{Log}_{x_1}(x)}.$$

Consider the first term. Differentiating the map $x \mapsto \frac{1}{1-t}\operatorname{Log}_{x_1}(x)$ yields a factor $\frac{1}{1-t}(d\operatorname{Log}_{x_1})_x$. Thus,

$$\begin{aligned}& \left\| \operatorname{Grad}_x \log\det (d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} \right\| \\ \leq & \frac{1}{1-t} \|(d\operatorname{Log}_{x_1})_x\| \left\| \operatorname{Grad}_\xi \log\det (d\operatorname{Exp}_{x_1})_\xi \right\| \Big|_{\xi = \frac{1}{1-t}\operatorname{Log}_{x_1}(x)}.\end{aligned}$$

Now apply Lemma 60 (specifically, the gradient bound in (34)) at $A(\xi) = (d\operatorname{Exp}_{x_1})_\xi$ to get

$$\left\| \operatorname{Grad}_\xi \log\det (d\operatorname{Exp}_{x_1})_\xi \right\| \leq d \, \|(d\operatorname{Exp}_{x_1})_\xi^{-1}\| \, \|\nabla (d\operatorname{Exp}_{x_1})_\xi\|.$$

Combining the above, we obtain

$$\begin{aligned}& \left\| \operatorname{Grad}_x \log\det (d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} \right\| \\ \leq & \frac{d}{1-t} \|(d\operatorname{Log}_{x_1})_x\| \, \|(d\operatorname{Exp}_{x_1})^{-1}_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\| \, \|\nabla (d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\|.\end{aligned}$$

The same argument (without the factor $\frac{1}{1-t}$) yields

$$\left\| \operatorname{Grad}_x \log\det (d\operatorname{Exp}_{x_1})_{\operatorname{Log}_{x_1}(x)} \right\| \leq d \, \|(d\operatorname{Log}_{x_1})_x\| \, \|(d\operatorname{Exp}_{x_1})^{-1}_{\operatorname{Log}_{x_1}(x)}\| \, \|\nabla (d\operatorname{Exp}_{x_1})_{\operatorname{Log}_{x_1}(x)}\|.$$

Finally, apply the triangle inequality, we get

$$\begin{aligned}\|\operatorname{Grad}_x \log J_t(x \mid x_1)\| \leq d \, \|(d\operatorname{Log}_{x_1})_x\| \Bigg(& \frac{1}{1-t} \|(d\operatorname{Exp}_{x_1})^{-1}_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\| \, \|\nabla (d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\| \\ & + \|(d\operatorname{Exp}_{x_1})^{-1}_{\operatorname{Log}_{x_1}(x)}\| \, \|\nabla (d\operatorname{Exp}_{x_1})_{\operatorname{Log}_{x_1}(x)}\| \Bigg).\end{aligned}$$

Now we estimate the Hessian. Write $f(x) := \log\det (d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}$. Taking second derivative (with product rule) we have

$$\begin{aligned}\|\nabla_x^2 f(x)\|_{\mathrm{op}} \leq & \left\| \nabla_\xi^2 \log\det (d\operatorname{Exp}_{x_1})_\xi \right\|_{\mathrm{op}} \Big|_{\xi = \frac{1}{1-t}\operatorname{Log}_{x_1}(x)} \cdot \left\| \frac{1}{1-t}(d\operatorname{Log}_{x_1})_x \right\|^2 \\ & + \left\| \operatorname{Grad}_\xi \log\det (d\operatorname{Exp}_{x_1})_\xi \right\| \Big|_{\xi = \frac{1}{1-t}\operatorname{Log}_{x_1}(x)} \cdot \left\| \nabla \Big(\frac{1}{1-t}(d\operatorname{Log}_{x_1})_x \Big) \right\|.\end{aligned}$$

Thus we have (here ∇_ξ^2 represent taking derivative w.r.t. ξ)

$$
\begin{aligned}
&\|\nabla_x^2 \log J_t(x \mid x_1)\|_{\mathrm{op}} \\
\le&\|(d\operatorname{Log}_{x_1})_x\|^2 \left(\frac{1}{(1-t)^2} \left\|\nabla_\xi^2 \log\det(d\operatorname{Exp}_{x_1})_\xi\right\|_{\mathrm{op}}\Big|_{\xi=\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} + \left\|\nabla_\xi^2 \log\det(d\operatorname{Exp}_{x_1})_\xi\right\|_{\mathrm{op}}\Big|_{\xi=\operatorname{Log}_{x_1}(x)} \right) \\
&+ \|\nabla(d\operatorname{Log}_{x_1})_x\| \left(\frac{1}{1-t} \left\| \operatorname{Grad}_\xi \log\det(d\operatorname{Exp}_{x_1})_\xi\right\|\Big|_{\xi=\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} + \left\| \operatorname{Grad}_\xi \log\det(d\operatorname{Exp}_{x_1})_\xi\right\|\Big|_{\xi=\operatorname{Log}_{x_1}(x)} \right) \\
\le& d\,\|(d\operatorname{Log}_{x_1})_x\|^2 \Bigg[\frac{1}{(1-t)^2} \Big(\|(d\operatorname{Exp}_{x_1})^{-1}_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\|^2 \, \|\nabla(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\|^2 \\
&+ \|(d\operatorname{Exp}_{x_1})^{-1}_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\| \, \|\nabla^2(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\| \Big) \\
&+ \Big(\|(d\operatorname{Exp}_{x_1})^{-1}_{\operatorname{Log}_{x_1}(x)}\|^2 \, \|\nabla(d\operatorname{Exp}_{x_1})_{\operatorname{Log}_{x_1}(x)}\|^2 + \|(d\operatorname{Exp}_{x_1})^{-1}_{\operatorname{Log}_{x_1}(x)}\| \, \|\nabla^2(d\operatorname{Exp}_{x_1})_{\operatorname{Log}_{x_1}(x)}\| \Big) \Bigg] \\
&+ d\,\|\nabla(d\operatorname{Log}_{x_1})_x\| \left(\frac{1}{1-t} \, \|(d\operatorname{Exp}_{x_1})^{-1}_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\| \, \|\nabla(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\| \right. \\
&\left. + \|(d\operatorname{Exp}_{x_1})^{-1}_{\operatorname{Log}_{x_1}(x)}\| \, \|\nabla(d\operatorname{Exp}_{x_1})_{\operatorname{Log}_{x_1}(x)}\| \right),
\end{aligned}
$$

where in the first inequality we used triangle inequality, and in the second inequality we applied Lemma 60.

For time derivative, since x is fixed and only $\frac{1}{1-t}$ depends on t,

$$
\partial_t \log J_t(x \mid x_1) = \frac{d}{1-t} + \partial_t \Big(\log\det(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} \Big).
$$

By the chain rule,

$$
\partial_t \Big(\log\det(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} \Big) = D \log\det(d\operatorname{Exp}_{x_1})_\xi \Big|_{\xi=\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} \Big[\partial_t \Big(\frac{1}{1-t} \operatorname{Log}_{x_1}(x) \Big) \Big].
$$

But $\partial_t\big(\frac{1}{1-t}\operatorname{Log}_{x_1}(x)\big) = \frac{1}{(1-t)^2}\operatorname{Log}_{x_1}(x)$, hence

$$
\left| \partial_t \Big(\log\det(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} \Big) \right| \le \frac{1}{(1-t)^2} \, \|\operatorname{Log}_{x_1}(x)\| \, \left\| \operatorname{Grad}_\xi \log\det(d\operatorname{Exp}_{x_1})_\xi \right\| \Big|_{\xi=\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}.
$$

Apply Lemma 60 (gradient bound (34)) at $\xi = \frac{1}{1-t}\operatorname{Log}_{x_1}(x)$, we obtain

$$
|\partial_t \log J_t(x \mid x_1)| \le \frac{d}{1-t} + \frac{d}{(1-t)^2} \, \|\operatorname{Log}_{x_1}(x)\| \, \|(d\operatorname{Exp}_{x_1})^{-1}_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\| \, \|\nabla(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\|.
$$

For mixed derivative, recall

$$
\operatorname{Grad}_x \log J_t(x \mid x_1) = \operatorname{Grad}_x \log\det(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)} - \operatorname{Grad}_x \log\det(d\operatorname{Exp}_{x_1})_{\operatorname{Log}_{x_1}(x)}.
$$

Only the first term depends on t. Differentiate the first term: there are two contributions, one from differentiating the prefactor $\frac{1}{1-t}$ and one from differentiating $\operatorname{Grad}_\xi \log\det(d\operatorname{Exp}_{x_1})_\xi$ at $\xi = \frac{1}{1-t}\operatorname{Log}_{x_1}(x)$. This yields

$$
\begin{aligned}
&\|\partial_t \operatorname{Grad}_x \log\det(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\| \\
\le&\|(d\operatorname{Log}_{x_1})_x\| \left[\frac{1}{(1-t)^2} \left\| \operatorname{Grad}_\xi \log\det(d\operatorname{Exp}_{x_1})_\xi \right\| \right|_{\xi=\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}
\end{aligned}
$$

$$+ \frac{1}{1-t} \left\| \partial_t \Big(\operatorname{Grad}_\xi \log\det(d\operatorname{Exp}_{x_1})_\xi \Big) \right\| \Bigg|_{\xi = \frac{1}{1-t}\operatorname{Log}_{x_1}(x)} \Bigg].$$

Next, note that

$$\partial_t(\operatorname{Grad}_\xi \log\det(d\operatorname{Exp}_{x_1})_\xi) = \nabla^2_\xi \log\det(d\operatorname{Exp}_{x_1})_\xi \, [\partial_t \xi]$$

with $\partial_t \xi = \frac{1}{(1-t)^2} \operatorname{Log}_{x_1}(x)$. Hence

$$\left\| \partial_t \Big(\operatorname{Grad}_\xi \log\det(d\operatorname{Exp}_{x_1})_\xi \Big) \right\| \leq \frac{1}{(1-t)^2} \left\| \operatorname{Log}_{x_1}(x) \right\| \left\| \nabla^2_\xi \log\det(d\operatorname{Exp}_{x_1})_\xi \right\|_{\mathrm{op}}.$$

By Lemma 60, we obtain

$$\begin{aligned}
&\|\partial_t \operatorname{Grad}_x \log J_t(x \mid x_1)\| \\
\leq &\|(d\operatorname{Log}_{x_1})_x\| \Bigg[\frac{1}{(1-t)^2} \, d \, \|(d\operatorname{Exp}_{x_1})^{-1}_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\| \, \|\nabla(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\| \\
&+ \frac{1}{(1-t)^3} \left\| \operatorname{Log}_{x_1}(x) \right\| d \Big(\|(d\operatorname{Exp}_{x_1})^{-1}_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\|^2 \, \|\nabla(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\|^2 \\
&+ \|(d\operatorname{Exp}_{x_1})^{-1}_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\| \, \|\nabla^2(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\| \Big) \Bigg].
\end{aligned}$$

Finally, by [22, Theorem 3.12] with $r = \frac{\|\operatorname{Log}_{x_1}(x)\|}{1-t}$, and noting that on $\operatorname{SPD}(n)$ sectional curvature is upper bounded by zero, we have

$$\|(d\operatorname{Exp}_{x_1})^{-1}_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\| \leq 1.$$

Hence we get

$$\begin{aligned}
\|\operatorname{Grad}_x \log J_t(x \mid x_1)\| \leq &d \, \|(d\operatorname{Log}_{x_1})_x\| \left(\frac{1}{1-t} \|\nabla(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\| + \|\nabla(d\operatorname{Exp}_{x_1})_{\operatorname{Log}_{x_1}(x)}\| \right), \\
\|\nabla^2_x \log J_t(x \mid x_1)\|_{\mathrm{op}} \leq &d \, \|(d\operatorname{Log}_{x_1})_x\|^2 \\
&\left[\frac{1}{(1-t)^2} \Big(\|\nabla(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\|^2 + \|\nabla^2(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\| \Big) \right. \\
&\left. + \Big(\|\nabla(d\operatorname{Exp}_{x_1})_{\operatorname{Log}_{x_1}(x)}\|^2 + \|\nabla^2(d\operatorname{Exp}_{x_1})_{\operatorname{Log}_{x_1}(x)}\| \Big) \right] \\
&+ d \, \|\nabla(d\operatorname{Log}_{x_1})_x\| \left(\frac{1}{1-t} \|\nabla(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\| + \|\nabla(d\operatorname{Exp}_{x_1})_{\operatorname{Log}_{x_1}(x)}\| \right), \\
|\partial_t \log J_t(x \mid x_1)| \leq &\frac{d}{1-t} + \frac{d}{(1-t)^2} \left\| \operatorname{Log}_{x_1}(x) \right\| \|\nabla(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\|, \\
\|\partial_t \operatorname{Grad}_x \log J_t(x \mid x_1)\| \leq &\|(d\operatorname{Log}_{x_1})_x\| \left[\frac{d}{(1-t)^2} \|\nabla(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\| \right. \\
&+ \frac{d}{(1-t)^3} \left\| \operatorname{Log}_{x_1}(x) \right\| \\
&\left. \Big(\|\nabla(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\|^2 + \|\nabla^2(d\operatorname{Exp}_{x_1})_{\frac{1}{1-t}\operatorname{Log}_{x_1}(x)}\| \Big) \right],
\end{aligned}$$

completing the proof. □

E.4 Bounding $\left\|\left(\nabla^2 d\,\mathbf{Exp}_p\right)_{rv}\right\|_{\mathrm{op}}$

In this section, we bound the third derivative of Exp. We first recall some notation in [22], which allows us to write the third derivative as some ODE. Then we apply some comparison theorem to derive the upper bound. We remark that on a symmetric space, the (covariant) derivative of the curvature tensor is identically zero, see for example [21, Theorem 10.19]. The techniques in this section works for non-symmetric spaces, but as we are working on $\mathrm{SPD}\,(n)$, we assume M is a symmetric space, which would simplify our computation.

Recall that ∇ denotes the (Levi–Civita) connection on TM over M. Let ∇^{flat} be the flat connection on the vector space T_pM. Set $\mathrm{Exp}_p : U \subset T_pM \to M$ on a normal neighborhood of p.

We first introduce pullback connection. Instead of differentiating along a vector field on M, we can also differentiate along a vector field X on T_pM. Define the corresponding connection as

$$\left(\nabla^{\text{pullback}}_X Y\right)(u) = \nabla_{(d\,\mathrm{Exp}_p)_u(X(u))}\, Y,$$

In other words, given $u \in T_pM$, and we have $q := \mathrm{Exp}_p(u)$. Then differentiate along X in the tangent space via the pullback connection, evaluated at $u \in T_pM$, is equivalent to differentiate along $d\,\mathrm{Exp}_p(X)$ (which can be viewed as a vector field on M) evaluated at q.

Recall the pullback bundle $\mathrm{Exp}_p^*(TM)$ is a vector bundle over the base T_pM. Concretely,

$$\mathrm{Exp}_p^*(TM) := \{(u, v) : u \in T_pM, v \in TM \text{ with base point } \mathrm{Exp}_p(u)\}.$$

For example, given an element of the tensor product space $(\alpha, (u, v)) \in T_p^*M \otimes \mathrm{Exp}_p^*(TM)$, we have $(\alpha, (u, v))(w) = \alpha(w)v$, where $\alpha \in T_p^*M$ is a covector, so that $\alpha(w) \in \mathbb{R}$.

Let $\{e_i\}$ be an orthonormal basis for T_pM, and $\{e^i\}$ be the corresponding basis in the cotangent space T_p^*M. By definition,

$$\big(\sum_i e^i \otimes (u, (d\,\mathrm{Exp}_p)_u[e_i])\big)(w) := \sum_i e^i(w)(d\,\mathrm{Exp}_p)_u[e_i],$$

and by linearity

$$\sum_i e^i(w)(d\,\mathrm{Exp}_p)_u[e_i] = (d\,\mathrm{Exp}_p)_u[\sum_i e^i(w)e_i] = (d\,\mathrm{Exp}_p)_u(w).$$

Thus we see that $d\,\mathrm{Exp}_p$ can be viewed as an element of the tensor product space:

$$(d\,\mathrm{Exp}_p)_u \sum_i e^i \otimes (u, (d\,\mathrm{Exp}_p)_u[e_i]) \in T_p^*M \otimes \mathrm{Exp}_p^*(TM).$$

We equip T_p^*M with the connection induced by ∇^{flat} and $\mathrm{Exp}_p^*(TM)$ with ∇^{pullback}. These induce a connection on the tensor product bundle, denoted by ∇^{induced}, characterized by the Leibniz rule. Now consider $\overline{W}_1, \overline{W}_2$ being vector fields in the tangent space T_pM, and $d\,\mathrm{Exp}_p(\overline{W}_2)$ can be viewed as a section, i.e., $d\,\mathrm{Exp}_p(\overline{W}_2) : u \mapsto (d\,\mathrm{Exp}_p)_u(\overline{W}_2(u))$. We differentiate

$$\begin{aligned}
&\nabla^{\text{pullback}}_{\overline{W}_1}(d\,\mathrm{Exp}_p(\overline{W}_2)) = \nabla^{\text{pullback}}_{\overline{W}_1}(\sum_i e^i(\overline{W_2})(d\,\mathrm{Exp}_p)[e_i]) \\
=&\sum_i \overline{W}_1 e^i(\overline{W_2})(d\,\mathrm{Exp}_p)[e_i] + \sum_i e^i(\overline{W_2})\nabla^{\text{pullback}}_{\overline{W}_1}(d\,\mathrm{Exp}_p)[e_i].
\end{aligned}$$

where we emphasize that $(d\,\mathrm{Exp}_p)[e_i]$ is a function of u, similar as before: $(d\,\mathrm{Exp}_p)[e_i] : u \mapsto (d\,\mathrm{Exp}_p)_u[e_i]$.

By linearity of $(d\,\mathrm{Exp}_p)$, and definition of flat connection, we have

$$\sum_i \overline{W}_1 e^i(\overline{W_2})(d\,\mathrm{Exp}_p)[e_i] = (d\,\mathrm{Exp}_p)[\sum_i \overline{W}_1 e^i(\overline{W_2})e_i] = (d\,\mathrm{Exp}_p)[\nabla^{\text{flat}}_{\overline{W}_1}\overline{W}_2].$$

On the other hand, since e^i are constants, we know $\nabla^{\text{flat}}_{\overline{W}_1} e^i = 0$, so that

$$\begin{aligned} e^i \otimes \nabla^{\text{pullback}}_{\overline{W}_1}(d\operatorname{Exp}_p)[e_i] =& e^i \otimes \nabla^{\text{pullback}}_{\overline{W}_1}(d\operatorname{Exp}_p)[e_i] + \nabla^{\text{flat}}_{\overline{W}_1} e^i \otimes (d\operatorname{Exp}_p)[e_i] \\ =& \nabla^{\text{induced}}_{\overline{W}_1}(e^i \otimes (d\operatorname{Exp}_p)[e_i]), \end{aligned}$$

hence

$$\begin{aligned} &\sum_i e^i(\overline{W}_2) \otimes \nabla^{\text{pullback}}_{\overline{W}_1}((d\operatorname{Exp}_p)[e_i]) = \Big(\sum_i e^i \otimes \nabla^{\text{pullback}}_{\overline{W}_1}((d\operatorname{Exp}_p)[e_i])\Big)(\overline{W}_2) \\ =&\Big(\sum_i \nabla^{\text{induced}}_{\overline{W}_1}(e^i \otimes (d\operatorname{Exp}_p)[e_i])\Big)(\overline{W}_2) = \sum_i \Big(\nabla^{\text{induced}}_{\overline{W}_1}(e^i \otimes (d\operatorname{Exp}_p)[e_i])\Big)(\overline{W}_2) \\ =&(\nabla^{\text{induced}}_{\overline{W}_1}(\sum_i e^i \otimes (d\operatorname{Exp}_p)[e_i]))(\overline{W}_2) = (\nabla^{\text{induced}}_{\overline{W}_1} d\operatorname{Exp}_p)(\overline{W}_2). \end{aligned}$$

Substitute into the previous expression, we obtain

$$\begin{aligned} &\nabla^{\text{pullback}}_{\overline{W}_1}(d\operatorname{Exp}_p(\overline{W}_2)) \\ =&\sum_i \overline{W}_1 e^i(\overline{W}_2)(d\operatorname{Exp}_p)[e_i] + \sum_i e^i(\overline{W}_2)\nabla^{\text{pullback}}_{\overline{W}_1}(d\operatorname{Exp}_p)[e_i] \\ =&(d\operatorname{Exp}_p)[\nabla^{\text{flat}}_{\overline{W}_1} \overline{W}_2] + \nabla^{\text{induced}}_{\overline{W}_1}(d\operatorname{Exp}_p)(\overline{W}_2) \end{aligned}$$

Thus, for vector fields $\overline{W}_1, \overline{W}_2$ on $U \subset T_pM$ we have

$$\nabla^{\text{pullback}}_{\overline{W}_1}\big(d\operatorname{Exp}_p(\overline{W}_2)\big) = \big(\nabla^{\text{induced}}_{\overline{W}_1} d\operatorname{Exp}_p\big)(\overline{W}_2) + d\operatorname{Exp}_p\big(\nabla^{\text{flat}}_{\overline{W}_1} \overline{W}_2\big). \tag{26}$$

Define $c(t, s_1, s_2, s_3) = \operatorname{Exp}_p(t(v + s_1w_1 + s_2w_2 + s_3w_3))$, and $\gamma(t) = \operatorname{Exp}_p(tv)$. We briefly recall what [22] did, to control the second derivative of Exp map. Define $J_1(t) = (d\operatorname{Exp}_p)_{tv}(tw_1)$ to be the Jacobi field along γ with initial condition w_1. We can define the extension of J_1 in the w_2-direction by $\widetilde{J}_1(t, s) := (d\operatorname{Exp}_p)_{t(v+sw_2)}(tw_1)$, where note that $\widetilde{J}_1(t, 0) = J_1(t)$. For fixed t, the tangent space vector field $u \mapsto tw_1$ is constant on T_pM, hence

$$\nabla^{\text{flat}}_{tw_2}(tw_1) = 0. \tag{27}$$

Applying (26) with $\overline{W}_1 \equiv tw_2$ and $\overline{W}_2 \equiv tw_1$, and evaluated at $t(v + sw_2)$, we have

$$\big(\nabla^{\text{pullback}}_{tw_2} \widetilde{J}_1(t, s)\big)(t(v + sw_2)) = \big(\nabla^{\text{induced}}_{tw_2} d\operatorname{Exp}_p\big)_{t(v+sw_2)}(tw_1) + (d\operatorname{Exp}_p)_{t(v+sw_2)}\big(\nabla^{\text{flat}}_{tw_2}(tw_1)\big).$$

By (27) the last term vanishes. Using the fact $\big(\nabla^{\text{pullback}}_X Y\big)(u) = \nabla_{(d\operatorname{Exp}_p)_u(X(u))} Y$ with $X = tw_2$, $Y = \widetilde{J}_1$, $u = tv$, we have (restricted to $s = 0$)

$$\big(\nabla^{\text{pullback}}_{tw_2} \widetilde{J}_1\big)(tv) = \nabla_{(d\operatorname{Exp}_p)_{tv}(tw_2)} \widetilde{J}_1.$$

Together, with the previously defined notation $J_2 = (d\operatorname{Exp}_p)_{tv}(tw_2)$, we have

$$\nabla_{J_2} J_1 = \nabla_{J_2} \widetilde{J}_1\big|_{s=0} = \nabla^{\text{pullback}}_{tw_2} \widetilde{J}_1\big|_{s=0} = \big(\nabla^{\text{induced}} d\operatorname{Exp}_p\big)_{tv}(tw_1, tw_2).$$

We therefore define the second-order variation field along γ by

$$K_{12}(t) := \big(\nabla^{\text{induced}} d\operatorname{Exp}_p\big)_{tv}(tw_1, tw_2) = \nabla_{J_2} J_1.$$

Then [22, Proposition 4.1] shows that K_{12} satisfies

$$\ddot{K}_{12} + R(K_{12}, \dot{\gamma})\dot{\gamma} + Y_{12} = 0, \qquad K_{12}(0) = 0,\ \dot{K}_{12}(0) = 0, \tag{28}$$

with Y_{12} given explicitly therein. Using the same technique, we can define

$K_{13}(t) := \big(\nabla^{\text{induced}} d\operatorname{Exp}_p\big)_{tv}(tw_1, tw_3) = \nabla_{J_3} J_1, K_{23}(t) := \big(\nabla^{\text{induced}} d\operatorname{Exp}_p\big)_{tv}(tw_2, tw_3) = \nabla_{J_3} J_2,$

satisfying the corresponding ODE as (28).

Now we study the third derivative $L_{123} = \big((\nabla^{\text{induced}})^2 d\operatorname{Exp}_p\big)(tw_1, tw_2, tw_3)$. Define $J_i(t)$ to be the Jacobi field along γ with initial condition w_i for $i = 2, 3$, in the same way as we did for J_1. Extend $K_{12}(t) = \big(\nabla^{\text{induced}} d\operatorname{Exp}_p\big)_{tv}(tw_1, tw_2) = \nabla_{J_2} J_1$ off γ by

$$\widetilde{K}_{12}(t, s_3) := \big(\nabla^{\text{induced}} d\operatorname{Exp}_p\big)_{t(v+s_3 w_3)}(tw_1, tw_2),$$

so that $\widetilde{K}_{12}(t, 0) = K_{12}(t)$. We then define the third-order variation field along γ by

$$L_{123}(t) := \nabla_{J_3} K_{12}(t) := \nabla_{J_3} \widetilde{K}_{12}(t, s_3)\Big|_{s_3=0}. \tag{29}$$

Apply $\nabla^{\text{pullback}}_{\overline{W}_3}$ to the first-order Leibniz rule (26). For vector fields $\overline{W}_1, \overline{W}_2, \overline{W}_3$ on T_pM we obtain, at any $u \in T_pM$,

$$\begin{aligned}\big(\nabla^{\text{pullback}}_{\overline{W}_3} \nabla^{\text{pullback}}_{\overline{W}_1} (d\operatorname{Exp}_p(\overline{W}_2))\big)(u) = {} & \big((\nabla^{\text{induced}}_{\overline{W}_3} \nabla^{\text{induced}}_{\overline{W}_1} d\operatorname{Exp}_p)_u\big)(\overline{W}_2(u)) \\ & + \big(\nabla^{\text{induced}}_{\overline{W}_1} d\operatorname{Exp}_p\big)_u\big((\nabla^{\text{flat}}_{\overline{W}_3} \overline{W}_2)(u)\big) \\ & + \big(\nabla^{\text{induced}}_{\overline{W}_3} d\operatorname{Exp}_p\big)_u\big((\nabla^{\text{flat}}_{\overline{W}_1} \overline{W}_2)(u)\big) \\ & + (d\operatorname{Exp}_p)_u\big((\nabla^{\text{flat}}_{\overline{W}_3} \nabla^{\text{flat}}_{\overline{W}_1} \overline{W}_2)(u)\big).\end{aligned} \tag{30}$$

Now specialize (30) to the constant vector fields $\overline{W}_1 \equiv tw_2$, $\overline{W}_2 \equiv tw_1$, $\overline{W}_3 \equiv tw_3$. Since $\nabla^{\text{flat}}_{tw_i}(tw_j) = 0$ for all i, j, the last three terms in (30) vanish. Evaluating at $u = tv$ yields

$$L_{123}(t) = \big(\nabla^{\text{pullback}}_{tw_3} \widetilde{K}_{12}\big)(tv) = \big((\nabla^{\text{induced}}_{tw_3} \nabla^{\text{induced}}_{tw_2} d\operatorname{Exp}_p)_{tv}\big)(tw_1),$$

where note that $(d\operatorname{Exp}_p)_{tv}(tw_3) = J_3(t)$.

By definition of the second covariant derivative tensor $(\nabla^{\text{induced}})^2 d\operatorname{Exp}_p$, this is precisely

$$L_{123}(t) = \big((\nabla^{\text{induced}})^2 d\operatorname{Exp}_p\big)_{tv}(tw_1, tw_2, tw_3). \tag{31}$$

Throughout this section, we consider covariant derivative along J_i, for example $\nabla_{J_3} J_2$, as abbreviation of $\nabla_{J_3} \tilde{J}_2$. Here, the $\tilde{\cdot}$ represent the corresponding extension along suitable direction, as constructed above.

Lemma 43. *For L_{123} defined above, we have*

$$\ddot{L}_{123} + R(L_{123}, \dot{\gamma})\dot{\gamma} + Y_{123} = 0, \qquad L_{123}(0) = 0,\ \dot{L}_{123}(0) = 0.$$

When M is a symmetric space,

$$\begin{aligned}Y_{123} = {} & R(\dot{J}_3, \dot{\gamma})K_{12} + 2R(J_3, \dot{\gamma})\dot{K}_{12} + R(K_{12}, \dot{J}_3)\dot{\gamma} + R(K_{12}, \dot{\gamma})\dot{J}_3 \\ & + 2R(\nabla_{J_3} J_1, \dot{\gamma})\dot{J}_2 + 2R(J_1, \dot{J}_3)\dot{J}_2 + 2R(J_1, \dot{\gamma})\nabla_{J_3}\dot{J}_2 \\ & + 2R(\nabla_{J_3} J_2, \dot{\gamma})\dot{J}_1 + 2R(J_2, \dot{J}_3)\dot{J}_1 + 2R(J_2, \dot{\gamma})\nabla_{J_3}\dot{J}_1.\end{aligned}$$

Proof. Now recall the following ODE is staisfied by K [22, Proposition 4.1.].

$$\ddot{K}_{12} + R(K_{12}, \dot{\gamma})\dot{\gamma} + Y_{12} = 0.$$

By differentiating the ODE (note that the vector field to which we take covariant derivative, is viewed as the corresponding extension, i.e., $\tilde{K}_{12}$ for K_{12}, Grad r for $\dot{\gamma}$ and $\nabla_{\text{Grad } r}\nabla_{\text{Grad } r}\tilde{K}_{12}$ for $\ddot{K}_{12}$), we obtain

$$\nabla_{J_3}\ddot{K}_{12} + \nabla_{J_3} R(K_{12}, \dot{\gamma})\dot{\gamma} + \nabla_{J_3} Y_{12} = 0,$$

We first compute $\nabla_{J_3}\ddot{K}_{12}$. Recall $\dot{X} := \nabla_{\dot{\gamma}} X$ and $\ddot{X} := \nabla_{\dot{\gamma}}\nabla_{\dot{\gamma}} X$. Let $L_{123} := \nabla_{J_3} K_{12}$. We compute $\nabla_{J_3}\ddot{K}_{12}$ step by step.

Recall by definition of curvature tensor

$$\nabla_{J_3}\nabla_{\dot{\gamma}} X = \nabla_{\dot{\gamma}}\nabla_{J_3} X + \nabla_{[J_3, \dot{\gamma}]} X + R(J_3, \dot{\gamma})X,$$

and the fact that $[J_3, \dot\gamma] = 0$ (for the same reason as in [22, Proposition 4.1]), we have

$$\begin{aligned}\nabla_{J_3}\ddot K_{12} &= \nabla_{J_3}\nabla_{\dot\gamma}\dot K_{12} = \nabla_{\dot\gamma}\nabla_{J_3}\dot K_{12} + R(J_3,\dot\gamma)\dot K_{12}\\ &= \nabla_{\dot\gamma}\nabla_{J_3}\nabla_{\dot\gamma}K_{12} + R(J_3,\dot\gamma)\dot K_{12}\\ &= \nabla_{\dot\gamma}(\nabla_{\dot\gamma}\nabla_{J_3}K_{12} + R(J_3,\dot\gamma)K_{12}) + R(J_3,\dot\gamma)\dot K_{12} = \nabla_{\dot\gamma}\nabla_{\dot\gamma}L_{123} + \nabla_{\dot\gamma}(R(J_3,\dot\gamma)K_{12}) + R(J_3,\dot\gamma)\dot K_{12}.\end{aligned}$$

Notice that (by Leibniz rule)

$$\begin{aligned}\nabla_{\dot\gamma}(R(J_3,\dot\gamma)K_{12}) &= (\nabla_{\dot\gamma}R)(J_3,\dot\gamma)K_{12} + R(\nabla_{\dot\gamma}J_3,\dot\gamma)K_{12} + R(J_3,\nabla_{\dot\gamma}\dot\gamma)K_{12} + R(J_3,\dot\gamma)\nabla_{\dot\gamma}K_{12}\\ &= (\nabla_{\dot\gamma}R)(J_3,\dot\gamma)K_{12} + R(\dot J_3,\dot\gamma)K_{12} + R(J_3,\dot\gamma)\dot K_{12}.\end{aligned}$$

Hence we have

$$\nabla_{J_3}\ddot K_{12} = \ddot L_{123} + (\nabla_{\dot\gamma}R)(J_3,\dot\gamma)K_{12} + R(\dot J_3,\dot\gamma)K_{12} + 2R(J_3,\dot\gamma)\dot K_{12}.$$

Now we compute $\nabla_{J_3}\big(R(K_{12},\dot\gamma)\dot\gamma\big)$. Note that $\nabla_{J_3}\dot\gamma = \nabla_{\dot\gamma}J_3 = \dot J_3$, by Leibniz rule,

$$\begin{aligned}\nabla_{J_3}\big(R(K_{12},\dot\gamma)\dot\gamma\big) &= (\nabla_{J_3}R)(K_{12},\dot\gamma)\dot\gamma + R(\nabla_{J_3}K_{12},\dot\gamma)\dot\gamma + R(K_{12},\nabla_{J_3}\dot\gamma)\dot\gamma + R(K_{12},\dot\gamma)\nabla_{J_3}\dot\gamma\\ &= (\nabla_{J_3}R)(K_{12},\dot\gamma)\dot\gamma + R(L_{123},\dot\gamma)\dot\gamma + R(K_{12},\dot J_3)\dot\gamma + R(K_{12},\dot\gamma)\dot J_3.\end{aligned}$$

Recall

$$Y := 2R(J_1,\dot\gamma)\dot J_2 + 2R(J_2,\dot\gamma)\dot J_1 + (\nabla_{\dot\gamma}R)(J_2,\dot\gamma)J_1 + (\nabla_{J_2}R)(J_1,\dot\gamma)\dot\gamma.$$

Also recall $[J_3,\dot\gamma] = 0$ and $\nabla_{J_3}\dot\gamma = \nabla_{\dot\gamma}J_3 = \dot J_3$. Then, by the Leibniz rule for tensor fields, we have

$$\begin{aligned}\nabla_{J_3}Y = {}& 2\,\nabla_{J_3}\big(R(J_1,\dot\gamma)\dot J_2\big) + 2\,\nabla_{J_3}\big(R(J_2,\dot\gamma)\dot J_1\big)\\ &+ \nabla_{J_3}\Big((\nabla_{\dot\gamma}R)(J_2,\dot\gamma)J_1\Big) + \nabla_{J_3}\Big((\nabla_{J_2}R)(J_1,\dot\gamma)\dot\gamma\Big),\end{aligned}$$

where each term expands as follows. For the first term,

$$\begin{aligned}\nabla_{J_3}\big(R(J_1,\dot\gamma)\dot J_2\big) &= (\nabla_{J_3}R)(J_1,\dot\gamma)\dot J_2 + R(\nabla_{J_3}J_1,\dot\gamma)\dot J_2 + R(J_1,\nabla_{J_3}\dot\gamma)\dot J_2 + R(J_1,\dot\gamma)\nabla_{J_3}\dot J_2\\ &= (\nabla_{J_3}R)(J_1,\dot\gamma)\dot J_2 + R(\nabla_{J_3}J_1,\dot\gamma)\dot J_2 + R(J_1,\dot J_3)\dot J_2 + R(J_1,\dot\gamma)\nabla_{J_3}\dot J_2.\end{aligned}$$

For the second term,

$$\begin{aligned}\nabla_{J_3}\big(R(J_2,\dot\gamma)\dot J_1\big) &= (\nabla_{J_3}R)(J_2,\dot\gamma)\dot J_1 + R(\nabla_{J_3}J_2,\dot\gamma)\dot J_1 + R(J_2,\nabla_{J_3}\dot\gamma)\dot J_1 + R(J_2,\dot\gamma)\nabla_{J_3}\dot J_1\\ &= (\nabla_{J_3}R)(J_2,\dot\gamma)\dot J_1 + R(\nabla_{J_3}J_2,\dot\gamma)\dot J_1 + R(J_2,\dot J_3)\dot J_1 + R(J_2,\dot\gamma)\nabla_{J_3}\dot J_1.\end{aligned}$$

For the third term,

$$\begin{aligned}\nabla_{J_3}\Big((\nabla_{\dot\gamma}R)(J_2,\dot\gamma)J_1\Big) &= (\nabla_{J_3}\nabla_{\dot\gamma}R)(J_2,\dot\gamma)J_1 + (\nabla_{\dot\gamma}R)(\nabla_{J_3}J_2,\dot\gamma)J_1\\ &\quad + (\nabla_{\dot\gamma}R)(J_2,\nabla_{J_3}\dot\gamma)J_1 + (\nabla_{\dot\gamma}R)(J_2,\dot\gamma)\nabla_{J_3}J_1\\ &= (\nabla_{J_3}\nabla_{\dot\gamma}R)(J_2,\dot\gamma)J_1 + (\nabla_{\dot\gamma}R)(\nabla_{J_3}J_2,\dot\gamma)J_1\\ &\quad + (\nabla_{\dot\gamma}R)(J_2,\dot J_3)J_1 + (\nabla_{\dot\gamma}R)(J_2,\dot\gamma)\nabla_{J_3}J_1.\end{aligned}$$

For the last term,

$$\begin{aligned}\nabla_{J_3}\Big((\nabla_{J_2}R)(J_1,\dot\gamma)\dot\gamma\Big) &= (\nabla_{J_3}\nabla_{J_2}R)(J_1,\dot\gamma)\dot\gamma + (\nabla_{J_2}R)(\nabla_{J_3}J_1,\dot\gamma)\dot\gamma\\ &\quad + (\nabla_{J_2}R)(J_1,\nabla_{J_3}\dot\gamma)\dot\gamma + (\nabla_{J_2}R)(J_1,\dot\gamma)\nabla_{J_3}\dot\gamma\\ &= (\nabla_{J_3}\nabla_{J_2}R)(J_1,\dot\gamma)\dot\gamma + (\nabla_{J_2}R)(\nabla_{J_3}J_1,\dot\gamma)\dot\gamma\\ &\quad + (\nabla_{J_2}R)(J_1,\dot J_3)\dot\gamma + (\nabla_{J_2}R)(J_1,\dot\gamma)\dot J_3.\end{aligned}$$

Plug in the expressions, we have

$$\ddot L_{123} + R(L_{123},\dot\gamma)\dot\gamma + Y_{123} = 0,$$

where

$$\begin{aligned}
Y_{123} &= (\nabla_{\dot\gamma}R)(J_3,\dot\gamma)K_{12} + (\nabla_{J_3}R)(K_{12},\dot\gamma)\dot\gamma + R(\dot J_3,\dot\gamma)K_{12} + 2R(J_3,\dot\gamma)\dot K_{12}\\
&\quad + R(K_{12},\dot J_3)\dot\gamma + R(K_{12},\dot\gamma)\dot J_3\\
&\quad + 2(\nabla_{J_3}R)(J_1,\dot\gamma)\dot J_2 + 2R(\nabla_{J_3}J_1,\dot\gamma)\dot J_2 + 2R(J_1,\dot J_3)\dot J_2 + 2R(J_1,\dot\gamma)\nabla_{J_3}\dot J_2\\
&\quad + 2(\nabla_{J_3}R)(J_2,\dot\gamma)\dot J_1 + 2R(\nabla_{J_3}J_2,\dot\gamma)\dot J_1 + 2R(J_2,\dot J_3)\dot J_1 + 2R(J_2,\dot\gamma)\nabla_{J_3}\dot J_1\\
&\quad + (\nabla_{J_3}\nabla_{\dot\gamma}R)(J_2,\dot\gamma)J_1 + (\nabla_{\dot\gamma}R)(\nabla_{J_3}J_2,\dot\gamma)J_1 + (\nabla_{\dot\gamma}R)(J_2,\dot J_3)J_1 + (\nabla_{\dot\gamma}R)(J_2,\dot\gamma)\nabla_{J_3}J_1\\
&\quad + (\nabla_{J_3}\nabla_{J_2}R)(J_1,\dot\gamma)\dot\gamma + (\nabla_{J_2}R)(\nabla_{J_3}J_1,\dot\gamma)\dot\gamma + (\nabla_{J_2}R)(J_1,\dot J_3)\dot\gamma + (\nabla_{J_2}R)(J_1,\dot\gamma)\dot J_3.
\end{aligned}$$

On a symmetric space,

$$\begin{aligned}
Y_{123} &= R(\dot J_3,\dot\gamma)K_{12} + 2R(J_3,\dot\gamma)\dot K_{12} + R(K_{12},\dot J_3)\dot\gamma + R(K_{12},\dot\gamma)\dot J_3\\
&\quad + 2R(\nabla_{J_3}J_1,\dot\gamma)\dot J_2 + 2R(J_1,\dot J_3)\dot J_2 + 2R(J_1,\dot\gamma)\nabla_{J_3}\dot J_2\\
&\quad + 2R(\nabla_{J_3}J_2,\dot\gamma)\dot J_1 + 2R(J_2,\dot J_3)\dot J_1 + 2R(J_2,\dot\gamma)\nabla_{J_3}\dot J_1.
\end{aligned}$$

□

Now we can apply comparison theory [22, Proposition 4.9] to control $\|L\|$.

Lemma 44. *Let (M,g) be a Riemannian* symmetric space*, i.e. $\nabla R\equiv 0$. Fix a unit-speed geodesic $\gamma:[0,r]\to M$ with $\|\dot\gamma\|\equiv 1$. Under Assumption 1, we have*

$$\|L(t)\| \le \frac{7L_R^2}{3}s_{K_{\min}}(s)^5 + 16L_R^3t^2s_{K_{\min}}(t)^5 + 20L_R^2ts_{K_{\min}}(t)^4.$$

Consequently,

$$\begin{aligned}
&\big\|\big((\nabla)^2d\,Exp_p\big)_{rv}\big\|_{\mathrm{op}} = \sup_{\|w_1\|,\|w_2\|,\|w_3\|\le 1}\big\|\big((\nabla^{induced})^2d\,Exp_p\big)_{rv}(w_1,w_2,w_3)\big\|\\
&= \sup_{\|w_1\|,\|w_2\|,\|w_3\|\le 1}\frac{1}{r^3}\|L_{123}(r)\| \le \frac{7L_R^2}{3r^3}s_{K_{\min}}(r)^5 + 16L_R^3\frac{1}{r}s_{K_{\min}}(r)^5 + 20L_R^2\frac{1}{r^2}s_{K_{\min}}(r)^4.
\end{aligned}$$

Proof. Recall

$$\begin{aligned}
Y_{123} &= R(\dot J_3,\dot\gamma)K_{12} + 2R(J_3,\dot\gamma)\dot K_{12} + R(K_{12},\dot J_3)\dot\gamma + R(K_{12},\dot\gamma)\dot J_3\\
&\quad + 2R(\nabla_{J_3}J_1,\dot\gamma)\dot J_2 + 2R(J_1,\dot J_3)\dot J_2 + 2R(J_1,\dot\gamma)\nabla_{J_3}\dot J_2\\
&\quad + 2R(\nabla_{J_3}J_2,\dot\gamma)\dot J_1 + 2R(J_2,\dot J_3)\dot J_1 + 2R(J_2,\dot\gamma)\nabla_{J_3}\dot J_1.
\end{aligned}$$

By definition of K_{13},K_{23}, we have $\nabla_{J_3}J_1 = K_{13}$ and $\nabla_{J_3}J_2 = K_{23}$. Next we express $\nabla_{J_3}\dot J_i$ in terms of $\dot K_{i3}$ and curvature. By definition of R and $[J_3,\dot\gamma]=0$,

$$\begin{aligned}
\nabla_{J_3}\dot J_i = \nabla_{J_3}\nabla_{\dot\gamma}J_i &= \nabla_{\dot\gamma}\nabla_{J_3}J_i + R(J_3,\dot\gamma)J_i\\
&= \dot K_{i3} + R(J_3,\dot\gamma)J_i.
\end{aligned}$$

Therefore

$$\begin{aligned}
2R(J_1,\dot\gamma)\nabla_{J_3}\dot J_2 &= 2R(J_1,\dot\gamma)\dot K_{23} + 2R(J_1,\dot\gamma)\big(R(J_3,\dot\gamma)J_2\big),\\
2R(J_2,\dot\gamma)\nabla_{J_3}\dot J_1 &= 2R(J_2,\dot\gamma)\dot K_{13} + 2R(J_2,\dot\gamma)\big(R(J_3,\dot\gamma)J_1\big).
\end{aligned}$$

Substituting these and using $R(x,y)z \le L_R\|x\|\|y\|\|z\|$, we have

$$\begin{aligned}
\|Y_{123}\| &\le 3L_R\|K_{12}\|\|\dot J_3\| + 2L_R\|J_3\|\|\dot K_{12}\| + 2L_R\|K_{13}\|\|\dot J_2\| + 2L_R\|K_{23}\|\|\dot J_1\|\\
&\quad + 2L_R\|J_1\|\|\dot K_{23}\| + 2L_R\|J_2\|\|\dot K_{13}\| + 2L_R\|J_1\|\|\dot J_2\|\|\dot J_3\| + 2L_R\|J_2\|\|\dot J_1\|\|\dot J_3\|\\
&\quad + 4L_R^2\|J_1\|\|J_2\|\|J_3\|\\
&\le 7L_R\|K\|\|\dot J\| + 6L_R\|J\|\|\dot K\| + 4L_R\|J\|\|\dot J\|^2 + 4L_R^2\|J\|^3.
\end{aligned}$$

Using Lemma 45, [22, Theorem 3.12, Theorem 4.11] we have

$$\begin{aligned}\|K(t)\| &\le \frac{16}{3} s_{K_{\min}}(t/2)^2 L_R s_{K_{\min}}(t),\\ \|\dot{K}(t)\| &\le L_R^2 \frac{16}{3} t s_{K_{\min}}(t)^3 + 2L_R s_{K_{\min}}(t)^2,\\ \|J(t)\| &\le s_{K_{\min}}(t),\\ \|\dot{J}(t)\| &\le s'_{K_{\min}}(t),\end{aligned}$$

hence we can bound

$$\begin{aligned}\|Y_{123}\| \le& 7L_R \frac{16}{3} s_{K_{\min}}(t/2)^2 L_R s_{K_{\min}}(t) s'_{K_{\min}}(t) + 6L_R s_{K_{\min}}(t)(L_R^2 \frac{16}{3} t s_{K_{\min}}(t)^3 + 2L_R s_{K_{\min}}(t)^2)\\ &+ 4L_R s_{K_{\min}}(t)(s'_{K_{\min}}(t))^2 + 4L_R^2 (s_{K_{\min}}(t))^3\\ \le& \frac{112L_R^2}{3} s_{K_{\min}}(t/2)^2 s_{K_{\min}}(t) s'_{K_{\min}}(t) + 32L_R^3 t s_{K_{\min}}(t)^4 + 20L_R^2 s_{K_{\min}}(t)^3\\ \le& \frac{28L_R^2}{3} s_{K_{\min}}(t)^3 s'_{K_{\min}}(t) + 32L_R^3 t s_{K_{\min}}(t)^4 + 20L_R^2 s_{K_{\min}}(t)^3.\end{aligned}$$

We apply [22, Proposition 4.9]. Consider ODE

$$\rho''(t) + K_{\min}\rho(t) = \eta(t).$$

Define $y(t) = s_{K_{\min}}(t)$. Note that y solves $y''(t) + K_{\min} y(t) = 0$. Then the function $\rho(t) = \int_0^t y(t-s)\eta(s)ds$ satisfies $\rho''(t) + K_{\min}\rho(t) = \eta(t)$. We apply this to each part of the bound on Y_{123}:

$$\begin{aligned}\int_0^t s_{K_{\min}}(t-s)\frac{28L_R^2}{3} s_{K_{\min}}(s)^3 s'_{K_{\min}}(s)ds &\le \frac{7L_R^2}{3} s_{K_{\min}}(t)\int_0^t 4 s_{K_{\min}}(s)^3 s'_{K_{\min}}(s)ds = \frac{7L_R^2}{3} s_{K_{\min}}(s)^5,\\ \int_0^t s_{K_{\min}}(t-s)32L_R^3 s_{K_{\min}}(s)^4 sds &\le 32L_R^3 s_{K_{\min}}(t)\int_0^t s_{K_{\min}}(s)^4 sds\\ &\le 32L_R^3 s_{K_{\min}}(t)^5 \int_0^t sds = 16L_R^3 t^2 s_{K_{\min}}(t)^5,\\ \int_0^t s_{K_{\min}}(t-s)20L_R^2 s_{K_{\min}}(s)^3 ds &\le 20L_R^2 t s_{K_{\min}}(t)^4.\end{aligned}$$

Hence we can use the following $\tilde{\rho}$ as upper bound for ρ:

$$\rho(t) \le \tilde{\rho}(t) := \frac{7L_R^2}{3} s_{K_{\min}}(s)^5 + 16L_R^3 t^2 s_{K_{\min}}(t)^5 + 20L_R^2 t s_{K_{\min}}(t)^4.$$

In other words,

$$\|L(t)\| \le \frac{7L_R^2}{3} s_{K_{\min}}(s)^5 + 16L_R^3 t^2 s_{K_{\min}}(t)^5 + 20L_R^2 t s_{K_{\min}}(t)^4.$$

□

Lemma 45. *When M is a symmetric space, we have*

$$\|\dot{K}(t)\| \le L_R^2 \frac{16}{3} t s_{K_{\min}}(t)^3 + 2L_R s_{K_{\min}}(t)^2.$$

Proof.

Since K solves $\ddot{K} + R(K,\dot{\gamma})\dot{\gamma} + Y = 0$, we have

$$\ddot{K} = -R(K,\dot{\gamma})\dot{\gamma} - Y.$$

Integrate over $[0,t]$, we obtain

$$\dot{K}(t) - \dot{K}(0) = \int_0^t \ddot{K}(s)ds = \int_0^t -R(K,\dot{\gamma})\dot{\gamma}(s) - Y(s)ds.$$

Using $\|\dot\gamma\| = 1$ and $\dot K(0) = 0$, we obtain

$$\|\dot K(t)\| = \|\int_0^t \ddot K(s)ds\| \le \int_0^t \|R(K,\dot\gamma)\dot\gamma(s)\| + \|Y(s)\|ds \le \int_0^t L_R\|K(s)\| + \|Y(s)\|ds.$$

It suffices to obtain bound for $\|Y\|$ and $\|K\|$. Recall

$$Y := 2R(J_1,\dot\gamma)\dot J_2 + 2R(J_2,\dot\gamma)\dot J_1 + (\nabla_{\dot\gamma}R)(J_2,\dot\gamma)J_1 + (\nabla_{J_2}R)(J_1,\dot\gamma)\dot\gamma,$$

and that for a symmetric space, we have Recall

$$Y = 2R(J_1,\dot\gamma)\dot J_2 + 2R(J_2,\dot\gamma)\dot J_1.$$

Hence

$$\|Y(s)\| \le 2L_R(\|J_1\|\|\dot J_2\| + \|\dot J_1\|\|J_2\|) \le 4L_R s'_{K_{\min}}(s)s_{K_{\min}}(s) = 2L_R s_{K_{\min}}(2s),$$

where by [22, Theorem 3.12] we have for unit tangent vector v,

$$\|d(\mathrm{Exp}_p)_{rv}(rw)\| \le \max\{1, \frac{s_{K_{\min}}(r)}{r}\}r\|w\|,$$

so that $\|J(s)\| \le s_{K_{\min}}(s)$. Also, [22, Theorem 3.11] gives

$$\|\operatorname{Hess} r\| \le \frac{s'_{K_{\min}}(r)}{s_{K_{\min}}(r)},$$

so that $\|\dot J(s)\| \le \|J\|\|\operatorname{Hess} r\| \le \frac{s'_{K_{\min}}(s)}{s_{K_{\min}}(s)}s_{K_{\min}}(s) \le s'_{K_{\min}}(s)$.

Furthermore, [22, Theorem 4.11] implies (with w replaced by rw)

$$K(s) \le \frac{16}{3}s_{K_{\min}}(s/2)^2 L_R s_{K_{\min}}(s).$$

Hence, we have

$$\|\dot K(t)\| \le \int_0^t L_R(\frac{16}{3}s_{K_{\min}}(s/2)^2 L_R s_{K_{\min}}(s)) + 2(L_R s_{K_{\min}}(2s))ds.$$

Note that in our case, $s_{K_{\min}}(t) = \frac{\sinh(\sqrt{|K_{\min}|}t)}{\sqrt{|K_{\min}|}}$, so we have $s_{K_{\min}}(s/2) \le s_{K_{\min}}(s)$.

For the first term, we have

$$\begin{aligned}\int_0^t L_R(\frac{16}{3}s_{K_{\min}}(s/2)^2 L_R s_{K_{\min}}(s))ds &\le L_R^2\frac{16}{3}\int_0^t s_{K_{\min}}(s)^3 ds \le L_R^2\frac{16}{3}\int_0^t s_{K_{\min}}(t)^3 ds \\ &= L_R^2\frac{16}{3}ts_{K_{\min}}(t)^3.\end{aligned}$$

For the second term,

$$\int_0^t 2(L_R s_{K_{\min}}(2s))ds \le 2L_R s_{K_{\min}}(t)^2.$$

Put together, we obtain

$$\|\dot K(t)\| \le L_R^2\frac{16}{3}ts_{K_{\min}}(t)^3 + 2L_R s_{K_{\min}}(t)^2.$$

□

E.5 Bounds involving Riemannian Log

Here, we obtain bounds on $\|(d\operatorname{Log}_x)_y\|$, $\big\|\big(\nabla(d\operatorname{Log}_x)\big)_y\big\|$, $\|\nabla_x \operatorname{Log}_x(y)\|_{\mathrm{op}}$, $\|\nabla_x \operatorname{Log}_x(y)\|_{\mathrm{op}}^2$, $|\operatorname{Div}_x \operatorname{Log}_x(y)|$, and $|\operatorname{Div}_x \operatorname{Log}_x(y)|^2$. When the manifold is $\mathrm{SPD}(n)$, we can obtain a sharper bound for $\|\operatorname{Grad} \operatorname{Div}_x \operatorname{Log}_x(y)\|$, see Lemma 49. Bounding $\|\operatorname{Grad} \operatorname{Div}_x \operatorname{Log}_x(y)\|$ for general Hadamard manifold with $\nabla R \equiv 0$ is more complicated, we defer the proof to Appendix E.6. In this section, for different differential operators d and ∇, note that $\nabla \operatorname{Log}_x(x_1)$ is covariant derivative (viewing $\operatorname{Log}_x(x_1)$ as a vector field, with x_1 being fixed. On the other hand, $d\operatorname{Log}_{x_1}$ is considering fixed base point.

Before presenting our results, we also recall some facts. Let M be a Hadamard manifold of dimension d. Fix $x \in M$ and let $y \in M$, $y \neq x$. Then:

(i) $\|\operatorname{Log}_x(y)\| = d(x,y)$.

(ii) $\operatorname{Grad}_x\Big(\frac{1}{2}d(x,y)^2\Big) = -\operatorname{Log}_x(y)$.

(iii) $\operatorname{Div}_x \operatorname{Log}_x(y) = -\Delta_x\Big(\frac{1}{2}d(x,y)^2\Big)$.

Lemma 46 (Bounds for $d\operatorname{Log}_x$ and $\nabla(d\operatorname{Log}_x)$)**.** *Let M be a Hadamard manifold of dimension d. Fix $x \in M$. Then, we have*

$$\begin{aligned}\|(d\,\mathit{Log}_x)_y\| &\leq 1,\\ \big\|\big(\nabla(d\,\mathit{Log}_x)\big)_y\big\| &\leq \big\|\big(\nabla(d\,\mathit{Exp}_x)\big)_{\mathit{Log}_x(y)}\big\|.\end{aligned}$$

Proof. On a Hadamard manifold, $\operatorname{Exp}_x : T_xM \to M$ is a global diffeomorphism, hence $\operatorname{Log}_x = \operatorname{Exp}_x^{-1}$ is smooth on M. Let $u = \operatorname{Log}_x(y)$. Then $\operatorname{Exp}_x(u) = y$ and the inverse function theorem gives

$$(d\operatorname{Log}_x)_y = \big((d\operatorname{Exp}_x)_u\big)^{-1} = \big((d\operatorname{Exp}_x)_{\operatorname{Log}_x(y)}\big)^{-1}.$$

Take norm on both sides and recall $\big\|\big((d\operatorname{Exp}_x)_{\operatorname{Log}_x(y)}\big)^{-1}\big\| \leq 1$, we get the first bound.

Now we differentiate the identity

$$(d\operatorname{Exp}_x)_{\operatorname{Log}_x(\cdot)} \circ (d\operatorname{Log}_x)_{\cdot} = Id_{T_{\cdot}M}$$

covariantly at y in the direction $w_1 \in T_yM$, and then apply the resulting operator to $w_2 \in T_yM$. Using Leibniz rule and that $\nabla(Id) = 0$, we get

$$\big(\nabla(d\operatorname{Exp}_x)\big)_{\operatorname{Log}_x(y)}\Big[(d\operatorname{Log}_x)_y w_1,\ (d\operatorname{Log}_x)_y w_2\Big] + (d\operatorname{Exp}_x)_{\operatorname{Log}_x(y)}\Big(\big(\nabla(d\operatorname{Log}_x)\big)_y[w_1,w_2]\Big) = 0.$$

Re-arrange the terms, we get

$$\begin{aligned}&\big(\nabla(d\operatorname{Log}_x)\big)_y[w_1,w_2]\\ &= -\big((d\operatorname{Exp}_x)_{\operatorname{Log}_x(y)}\big)^{-1}\Big(\big(\nabla(d\operatorname{Exp}_x)\big)_{\operatorname{Log}_x(y)}\Big[\big((d\operatorname{Exp}_x)_{\operatorname{Log}_x(y)}\big)^{-1}w_1,\ \big((d\operatorname{Exp}_x)_{\operatorname{Log}_x(y)}\big)^{-1}w_2\Big]\Big).\end{aligned}$$

Take norms, and recalling that $\|\big((d\operatorname{Exp}_x)_{\operatorname{Log}_x(y)}\big)^{-1}\| \leq 1$, we obtain

$$\big\|\big(\nabla(d\operatorname{Log}_x)\big)_y\big\| \leq \big\|\big(\nabla(d\operatorname{Exp}_x)\big)_{\operatorname{Log}_x(y)}\big\|.$$

□

Lemma 47 (Bounds for $\|\nabla_x \operatorname{Log}_x(y)\|$)**.** *Let M be a Hadamard manifold of dimension d with sectional curvature bounded below by $K_{\min} \leq 0$. Fix $x \in M$ and $y \in M$, $y \neq x$. Then*

$$\begin{aligned}\|\nabla_x\,\mathit{Log}_x(y)\|_{\mathrm{op}} &\leq 1 + d(x,y)\,\frac{s'_{K_{\min}}(d(x,y))}{s_{K_{\min}}(d(x,y))},\\ \|\nabla_x\,\mathit{Log}_x(y)\|_{\mathrm{op}}^2 &\leq 2 + 2\,d(x,y)^2\Big(\frac{s'_{K_{\min}}(d(x,y))}{s_{K_{\min}}(d(x,y))}\Big)^2.\end{aligned}$$

Proof. Let $r(\cdot) := d(\cdot, y)$, so $r(x) = d(x, y)$. Since $r\ \mathrm{Grad}_x\, r = -\mathrm{Log}_x(y)$ (since $\mathrm{Grad}_x(\frac{1}{2}r^2) = r\,\mathrm{Grad}_x\, r$), we have

$$\mathrm{Log}_x(y) = -\,r(x)\ \mathrm{Grad}_x\, r(x).$$

Differentiate covariantly in x, we obtain

$$\nabla_x \mathrm{Log}_x(y) = -\nabla_x\big(r(x)\ \mathrm{Grad}_x\, r(x)\big) = -\big(\nabla_x r(x)\big) \otimes \mathrm{Grad}_x\, r(x)\ -\ r(x)\,\nabla_x \mathrm{Grad}_x\, r(x).$$

Taking operator norms and using $\|\,\mathrm{Grad}_x\, r(x)\| = 1$ gives

$$\|\nabla_x \mathrm{Log}_x(y)\|_{\mathrm{op}} \le \|\,\mathrm{Grad}_x\, r(x)\|^2 + r(x)\,\|\nabla\,\mathrm{Grad}_x\, r(x)\|_{\mathrm{op}} = 1 + d(x,y)\,\|\nabla\,\mathrm{Grad}_x\, r(x)\|_{\mathrm{op}}.$$

By Hessian comparison under $\mathrm{Sec} \ge K_{\min}$, we have

$$\|\nabla\,\mathrm{Grad}_x\, r(x)\|_{\mathrm{op}} \le \frac{s'_{K_{\min}}(d(x,y))}{s_{K_{\min}}(d(x,y))}.$$

Combining the above estimates yields the first inequality. The squared bound follows from $(a+b)^2 \le 2a^2 + 2b^2$ with $a = 1$ and $b = d(x,y)\,\frac{s'_{K_{\min}}(d(x,y))}{s_{K_{\min}}(d(x,y))}$. □

Lemma 48 (Bounds for $\mathrm{Div}_x \mathrm{Log}_x(y)$ and $|\,\mathrm{Div}_x \mathrm{Log}_x(y)|^2$)**.** *Let M be a Hadamard manifold of dimension d with sectional curvature bounded below by $K_{\min} \le 0$. Fix $x \in M$ and $y \in M$, $y \ne x$. Then*

$$|\,\mathit{Div}_x \mathit{Log}_x(y)| \le 1 + (d-1)\,d(x,y)\,\frac{s'_{K_{\min}}(d(x,y))}{s_{K_{\min}}(d(x,y))},$$
$$|\,\mathit{Div}_x \mathit{Log}_x(y)|^2 \le 2 + 2(d-1)^2\,d(x,y)^2\Big(\frac{s'_{K_{\min}}(d(x,y))}{s_{K_{\min}}(d(x,y))}\Big)^2.$$

Proof. Let $r(\cdot) := d(\cdot, y)$. Then we have $\mathrm{Div}_x \mathrm{Log}_x(y) = -\Delta_x\Big(\frac{1}{2}r(x)^2\Big)$. Using the product rule for Laplacian, we have $\Delta(\frac{1}{2}r^2) = \langle \mathrm{Grad}\ r, \mathrm{Grad}\ r\rangle + r\Delta r = 1 + r\Delta r$. For the upper bound, Laplacian comparison theorem under $\mathrm{Sec} \ge K_{\min}$ gives

$$\Delta r \le (d-1)\,\frac{s'_{K_{\min}}(d(x,y))}{s_{K_{\min}}(d(x,y))}.$$

Substitute into $|\,\mathrm{Div}_x \mathrm{Log}_x(y)| = \Delta_x(\frac{1}{2}r^2)$ we obtain

$$|\,\mathrm{Div}_x \mathrm{Log}_x(y)| \le 1 + (d-1)\,d(x,y)\,\frac{s'_{K_{\min}}(d(x,y))}{s_{K_{\min}}(d(x,y))}.$$

Finally, using $(a+b)^2 \le 2a^2 + 2b^2$ with $a = 1$ and $b = (d-1)\,d(x,y)\,\frac{s'_{K_{\min}}(d(x,y))}{s_{K_{\min}}(d(x,y))}$ gives the desired results. □

Moreover, on $\mathrm{SPD}\,(n)$, we have the following results.

Lemma 49 ($\|\,\mathrm{Grad\ Div\ Log}\,\|$ bound on $\mathrm{SPD}\,(n)$)**.** *Let $\mathrm{SPD}\,(n)$ be the manifold of real symmetric positive-definite matrices endowed with the affine-invariant Riemannian metric*

$$\langle U, V\rangle_P := \mathrm{Tr}\,\big(P^{-1}UP^{-1}V\big), \qquad U, V \in T_P\,\mathrm{SPD}\,(n) = \mathit{Sym}(n).$$

Fix $y \in \mathrm{SPD}\,(n)$ and define

$$r(x) := d(x,y), \qquad f(x) := d(x,y)^2 = r(x)^2, \qquad d := \dim(\mathrm{SPD}\,(n)) = \frac{n(n+1)}{2}.$$

Then we have

$$\|\,\mathit{Grad\ Div\ Log}\,\| \le \frac{\sqrt{2}d}{2}\left(2\Big(1 + r\,\frac{s'_{K_{\min}}(r)}{s_{K_{\min}}(r)}\Big)\right)^{\frac{3}{2}}.$$

Proof. We first show that $\mathrm{SPD}\,(n)$ is a totally geodesic submanifold of $PD(n,\mathbb{C})$ under the affine-invariant metric. Let $\mathrm{PD}\,(n,\mathbb{C})$ denote the manifold of complex Hermitian positive-definite matrices endowed with the same affine-invariant metric $\langle U,V\rangle_P = \mathrm{Tr}\,(P^{-1}UP^{-1}V)$. Define a smooth isometry by

$$\sigma : \mathrm{PD}\,(n,\mathbb{C}) \to \mathrm{PD}\,(n,\mathbb{C}), \qquad \sigma(P) := \overline{P}.$$

Notice that σ is an isometry because for any $P \in \mathrm{PD}\,(n,\mathbb{C})$ and $U,V \in Herm(n)$,

$$\langle d\sigma_P[U], d\sigma_P[V]\rangle_{\sigma(P)} = \mathrm{Tr}\,\left(\overline{P}^{-1}\,\overline{U}\,\overline{P}^{-1}\,\overline{V}\right) = \overline{\mathrm{Tr}\,(P^{-1}UP^{-1}V)} = \mathrm{Tr}\,(P^{-1}UP^{-1}V) = \langle U,V\rangle_P,$$

where we used that $\mathrm{Tr}\,(P^{-1}UP^{-1}V) \in \mathbb{R}$ for Hermitian P,U,V. Thus we see that the fixed-point set of σ is exactly $\mathrm{SPD}\,(n)$. According to [17, Theorem 5.1] we have that the fixed-point set of an isometry is a (embedded) totally geodesic submanifold. Hence $\mathrm{SPD}\,(n)$ is totally geodesic in $\mathrm{PD}\,(n,\mathbb{C})$.

Let $F(P) := d_{\mathrm{PD}\,(n,\mathbb{C})}(P,y)^2$ be the squared distance to $y \in \mathrm{SPD}\,(n) \subset PD(n,\mathbb{C})$, considered as a function on $PD(n,\mathbb{C})$. By [14, Theorem 1.4], F is 2-self-concordant on $\mathrm{PD}\,(n,\mathbb{C})$: we have that for all $P \in \mathrm{PD}\,(n,\mathbb{C})$ and $u,v,w \in T_P\,\mathrm{PD}\,(n,\mathbb{C})$,

$$|(\nabla^3 F)_P(u,v,w)| \le \sqrt{2}\,\sqrt{(\nabla^2 F)_P(u,u)}\,\sqrt{(\nabla^2 F)_P(v,v)}\,\sqrt{(\nabla^2 F)_P(w,w)}.$$

Now restrict F to the totally geodesic submanifold $\mathrm{SPD}\,(n)$. By definition, for a totally geodesic submanifold, the second fundamental form vanishes, and Gauss formula [21, Theorem 8.2] implies that the connection ∇ on $\mathrm{SPD}\,(n)$ is induced by that of $\mathrm{PD}\,(n,\mathbb{C})$. Moreover, the geodesic distance function on $\mathrm{SPD}\,(n)$ is induced by that on $\mathrm{PD}\,(n,\mathbb{C})$, so we know self-concordance of distance squared also holds on $\mathrm{SPD}\,(n)$.

Now we make use of the self-concordance property to bound $\|\,\mathrm{Grad\ Div\ Log}\,\|$. Let $f(x) = d_{\mathrm{SPD}}\,(x,y)^2$ and note that on the Hadamard manifold $\mathrm{SPD}\,(n)$,

$$\|\,\mathrm{Grad}_x\,\mathrm{Div}_x\,\mathrm{Log}_x(y)\| = \tfrac{1}{2}\|\,\mathrm{Grad}_x\,\Delta_x f(x)\|.$$

Let $\{e_i\}_{i=1}^d$ be an orthonormal basis of $T_x\,\mathrm{SPD}\,(n)$. Such an orthonormal basis can be extended to a neighborhood of x, and we obtain an orthonormal frame: $\nabla e_i|_x = 0, \forall i$ [21, Exercise 5-21]. Thus we can differentiate Δf as follows. For any unit $w \in T_x\,\mathrm{SPD}\,(n)$,

$$\langle \mathrm{Grad}\ \Delta f, w\rangle = \sum_{i=1}^d (\nabla^3 f)(w,e_i,e_i) + 2\nabla^2 f(\nabla_w e_i, e_i) = \sum_{i=1}^d (\nabla^3 f)(w,e_i,e_i).$$

By self-concordance, we have $|(\nabla^3 f)(w,e_i,e_i)| \le \sqrt{2}\,\sqrt{(\nabla^2 f)(w,w)}\,(\nabla^2 f)(e_i,e_i)$. Therefore

$$\|\,\mathrm{Grad\ Div\ Log}\,\| \le \frac{\sqrt{2}d}{2}\,\sqrt{\|\nabla^2 f\|_{\mathrm{op}}^3}.$$

Writing $f = r^2$, we have

$$\nabla^2 f = 2\,dr \otimes dr + 2r\,\nabla^2 r, \qquad \Delta f = 2 + 2r\,\Delta r,$$

using $\|\,\mathrm{Grad}\ r\| = 1$. Under $\mathrm{Sec} \ge K_{\min}$, Hessian comparison gives $\|\nabla^2 r\|_{\mathrm{op}} \le \frac{s'_{K_{\min}}(r)}{s_{K_{\min}}(r)}$, hence

$$\|\nabla^2 f\|_{\mathrm{op}} \le 2 + 2r\,\|\nabla^2 r\|_{\mathrm{op}} \le 2\Big(1 + r\,\frac{s'_{K_{\min}}(r)}{s_{K_{\min}}(r)}\Big).$$

Substituting this in the previous estimate yields

$$\|\,\mathrm{Grad\ Div\ Log}\,\| \le \frac{\sqrt{2}d}{2}\left(2\Big(1 + r\,\frac{s'_{K_{\min}}(r)}{s_{K_{\min}}(r)}\Big)\right)^{1.5}.$$

□

E.6 Control the gradient of Div Log

In this section, we prove our bound on Grad Div Log for Hadamard manifolds that satisfies $\nabla R \equiv 0$. Our main result is presented in Proposition 54. We first define some notations. Let $x, y \in M$, and $v = \frac{\operatorname{Log}_y(x)}{\|\operatorname{Log}_y(x)\|}$. Let γ be the unique unit-speed minimizing geodesic from y to x, with $\gamma(s) = \operatorname{Exp}_y(sv)$. Consider the tangent space T_yM, for any $z \in T_yM$, let J_z be the Jacobi field along γ with initial conditions $J_z(0) = 0, D_sJ_z(0) = z$.

Since parallel transport commutes with the covariant derivative, the Jacobi equation of J_z implies

$$\frac{d^2}{ds^2} P^y_{\gamma(s)} J_z(s) + P^y_{\gamma(s)} \big(R(J_z(s), \gamma'(s))\gamma'(s)\big) = 0.$$

Define the linear map $D(s) : T_yM \to T_yM$ by parallel transport the Jacobi field back to T_yM:

$$D(s)z := P^y_{\gamma(s)} J_z(s).$$

We first characterize the operator $D(s)$.

Lemma 50. *Let M be a Hadamard manifold with $\nabla R \equiv 0$. Then the transported Jacobi field satisfies the following "Jacobi equation":*

$$\frac{d^2}{ds^2}\big(D(s)z\big) + R(D(s)z, v)v = 0, \qquad D(0)z = 0, \qquad \frac{d}{ds}D(0)z = z.$$

Define for every $s > 0$,

$$U(s) := \big(\frac{d}{ds}D(s)\big)\, D(s)^{-1}.$$

Then for every $z \in T_yM$,

$$\frac{d}{ds}\big(U(s)z\big) + U(s)\big(U(s)z\big) + R(z, v)v = 0.$$

Proof. Recall that

$$\frac{d^2}{ds^2} P^y_{\gamma(s)} J_z(s) + P^y_{\gamma(s)} \big(R(J_z(s), \gamma'(s))\gamma'(s)\big) = 0.$$

Notice that $J_z(s) = P_y^{\gamma(s)} D(s)z$ and $\gamma'(s) = P_y^{\gamma(s)} v$. Since $\nabla R = 0$, parallel transport preserves the curvature tensor along γ (This can be proved easily by Leibniz rule), hence

$$P^y_{\gamma(s)} \big(R(J_z(s), \gamma'(s))\gamma'(s)\big) = P^y_{\gamma(s)} \big(R(P_y^{\gamma(s)} D(s)z,\, P_y^{\gamma(s)} v)\, P_y^{\gamma(s)} v\big) = R(D(s)z, v)v.$$

Hence we have, for every $z \in T_yM$,

$$\frac{d^2}{ds^2}\big(D(s)z\big) + R(D(s)z, v)v = 0.$$

That is,

$$\frac{d^2}{ds^2}\big(D(s)z\big) + R(D(s)z, v)v = 0, \qquad D(0)z = 0, \qquad \frac{d}{ds}D(0)z = z. \tag{32}$$

We next derive the equation satisfied by $U(s) := \frac{d}{ds}D(s)\, D(s)^{-1}$. For every $z \in T_yM$,

$$U(s)z = \frac{d}{ds}D(s)\big(D(s)^{-1}z\big).$$

Differentiating with respect to s gives

$$\frac{d}{ds}\big(U(s)z\big) = \frac{d^2}{ds^2}D(s)\big(D(s)^{-1}z\big) + \frac{d}{ds}D(s)\left(\frac{d}{ds}D(s)^{-1}z\right).$$

Since

$$\frac{d}{ds}D(s)^{-1} = -D(s)^{-1}\frac{d}{ds}D(s)D(s)^{-1},$$

we obtain

$$\frac{d}{ds}\big(U(s)z\big) = \frac{d^2}{ds^2}D(s)\big(D(s)^{-1}z\big) - \frac{d}{ds}D(s)D(s)^{-1}\frac{d}{ds}D(s)D(s)^{-1}z.$$

Using the ODE for $D(s)$ with $D(s)^{-1}z$ in place of z,

$$\frac{d^2}{ds^2}D(s)\big(D(s)^{-1}z\big) = -R\big(D(s)D(s)^{-1}z, v\big)v = -R(z,v)v.$$

Hence

$$\frac{d}{ds}\big(U(s)z\big) = -R(z,v)v - U(s)\big(U(s)z\big).$$

Equivalently, for every $z \in T_yM$,

$$\frac{d}{ds}\big(U(s)z\big) + U(s)\big(U(s)z\big) + R(z,v)v = 0.$$

□

Lemma 51. *Recall that $\gamma'(0) = v$. We have that*

$$U(s)v = \big(\frac{d}{ds}D(s)\big)\big(D(s)^{-1}v\big) = \frac{1}{s}v.$$

Proof. We now express $\Delta d(\cdot, y)(\gamma(s))$ in terms of $U(s)$. First, we compute the radial direction. Consider the variation

$$\alpha(\tau, s) = \gamma((1+\tau)s).$$

For each fixed τ, the curve $s \mapsto \alpha(\tau, s)$ is a geodesic, so the variational vector field

$$W(s) := \frac{\partial}{\partial\tau}\alpha(\tau, s)\Big|_{\tau=0} = s\gamma'(s)$$

is a Jacobi field along γ. Also,

$$W(0) = 0, \qquad D_sW(0) = \gamma'(0) = v.$$

By uniqueness of Jacobi fields with prescribed initial data,

$$J_v(s) = s\gamma'(s).$$

Applying $P^y_{\gamma(s)}$ to both sides yields

$$D(s)v = P^y_{\gamma(s)}J_v(s) = P^y_{\gamma(s)}\big(s\gamma'(s)\big) = sv.$$

Differentiating gives

$$\frac{d}{ds}D(s)v = v,$$

and therefore

$$U(s)v = \big(\frac{d}{ds}D(s)\big)\big(D(s)^{-1}v\big) = \big(\frac{d}{ds}D(s)\big)\frac{1}{s}v = \frac{1}{s}v.$$

□

Lemma 52. *The operator $\frac{d}{dt}D(t)D(t)^{-1}$ can be viewed as the Hessian (of the distance function) on the subspace $v^\perp$ up to parallel transport, thus is self-adjoint. Moreover, the eigenvalues of $\frac{d}{dt}D(t)D(t)^{-1}$ on T_yM are the $d-1$ tangential Hessian eigenvalues and the additional radial eigenvalue $\frac{1}{t}$.*

Proof. Let $r(\cdot) := d(\cdot, y)$. For $t > 0$, the tangent space of the geodesic sphere of radius t centered at y at the point $\gamma(t)$ is

$$T_{\gamma(t)}S_t(y) = \gamma'(t)^\perp.$$

Define the linear operator $S_t : \gamma'(t)^\perp \to \gamma'(t)^\perp$ by

$$\langle S_t X, Y\rangle = \nabla^2 r\big|_{\gamma(t)}(X,Y) = \langle \nabla_X \text{ Grad } r, Y\rangle, \forall X, Y \in \gamma'(t)^\perp.$$

Since the Hessian is symmetric, the operator S_t is self-adjoint.

We now transport this operator back to T_yM. Define $\widetilde{S}_t : v^\perp \to v^\perp$ by

$$\widetilde{S}_t z := P^y_{\gamma(t)}\left(S_t\big(P^{\gamma(t)}_y z\big)\right), \qquad z \in v^\perp.$$

We show that

$$\widetilde{S}_t = \big(\frac{d}{dt}D(t)\big)D(t)^{-1} \qquad \text{on } v^\perp.$$

Let $z \in v^\perp$, and $u := D(t)^{-1}z$. Since $D(t)$ preserves $v^\perp$, we have $u \in v^\perp$. Also, $D(t)u = z$. By definition of $D(t)$,

$$P^y_{\gamma(t)} J_u(t) = D(t)u = z,$$

hence $J_u(t) = P^{\gamma(t)}_y z$. Therefore,

$$\widetilde{S}_t z = P^y_{\gamma(t)}\left(S_t\big(P^{\gamma(t)}_y z\big)\right) = P^y_{\gamma(t)}\left(S_t\big(J_u(t)\big)\right).$$

Since S_t is defined by the Hessian of r on $\gamma'(t)^\perp$,

$$S_t\big(J_u(t)\big) = \nabla_{J_u(t)} \text{ Grad } r.$$

We show that $\nabla_{J_u(t)} \text{ Grad } r = D_t J_u(t)$. Choose a smooth curve $w : (-\varepsilon, \varepsilon) \to T_yM$ such that $\|w(\tau)\| = 1$, $w(0) = v$, $w'(0) = u$. Define the variation

$$\Gamma(\tau, s) := \text{Exp}_y\big(s\, w(\tau)\big).$$

For each fixed τ, the curve $s \mapsto \Gamma(\tau, s)$ is a unit-speed geodesic issuing from y. Moreover, the variational vector field $\frac{\partial}{\partial\tau}\Gamma(\tau,s)\big|_{\tau=0}$ is the Jacobi field $J_u(s)$, since it satisfies $\frac{\partial}{\partial\tau}\Gamma(\tau,0)\big|_{\tau=0} = 0$ and $D_s\frac{\partial}{\partial\tau}\Gamma(\tau,s)\big|_{\tau=0,s=0} = u$. Applying the symmetry lemma ($D_\tau\partial_s\Gamma = D_s\partial_\tau\Gamma$), we obtain

$$\nabla_{J_u(t)} \text{ Grad } r = D_\tau \text{ Grad } r = D_\tau\left(\frac{\partial}{\partial s}\Gamma(\tau,s)\right)\Big|_{(\tau,s)=(0,t)} = D_s\left(\frac{\partial}{\partial\tau}\Gamma(\tau,s)\right)\Big|_{(\tau,s)=(0,t)} = D_t J_u(t).$$

Thus $\widetilde{S}_t z = P^y_{\gamma(t)}\big(D_t J_u(t)\big)$. Next, since $P^y_{\gamma(t)} J_u(t) = D(t)u$, differentiating with respect to t gives $P^y_{\gamma(t)}\big(D_t J_u(t)\big) = \frac{d}{dt}D(t)u$. Combine them together, we obtain

$$\widetilde{S}_t z = P^y_{\gamma(t)}\big(D_t J_u(t)\big) = \frac{d}{dt}D(t)u = \frac{d}{dt}D(t)D(t)^{-1}z.$$

Since this holds for every $z \in v^\perp$, we conclude that

$$\widetilde{S}_t = \frac{d}{dt}D(t)D(t)^{-1} \qquad \text{on } v^\perp.$$

Because $\widetilde{S}_t$ is obtained from the self-adjoint operator S_t by conjugation with the isometry $P^{\gamma(t)}_y$, it is also self-adjoint. Equivalently, for all

$$z, w \in v^\perp,$$

we have

$$\begin{aligned}\left\langle \frac{d}{dt}D(t)D(t)^{-1}z, w\right\rangle &= \left\langle \widetilde{S}_t z, w\right\rangle = \left\langle S_t\big(P^{\gamma(t)}_y z\big), P^{\gamma(t)}_y w\right\rangle = \left\langle P^{\gamma(t)}_y z, S_t\big(P^{\gamma(t)}_y w\big)\right\rangle \\ &= \left\langle z, \widetilde{S}_t w\right\rangle\end{aligned}$$

$$= \left\langle z, \frac{d}{dt} D(t) D(t)^{-1} w \right\rangle.$$

Therefore

$$\frac{d}{dt} D(t) D(t)^{-1}$$

is self-adjoint on $v^\perp$.

Since

$$\frac{d}{dt} D(t) D(t)^{-1} = P^y_{\gamma(t)} \circ S_t \circ P^{\gamma(t)}_y \qquad \text{on } v^\perp,$$

with S_t the self-adjoint operator representing $\nabla^2 d(\cdot, y)\big|_{\gamma(t)}$ on $\gamma'(t)^\perp$, the eigenvalues of $\frac{d}{dt} D(t) D(t)^{-1}$ on $v^\perp$ coincide with the eigenvalues of the Hessian of the distance function on $\gamma'(t)^\perp$. Together with

$$D(t)v = t\, v, \qquad \frac{d}{dt} D(t) D(t)^{-1} v = \frac{1}{t} v,$$

this shows that the eigenvalues of $\frac{d}{dt} D(t) D(t)^{-1}$ on $T_y M$ are the $d-1$ tangential Hessian eigenvalues and the additional radial eigenvalue $\frac{1}{t}$.

□

Now we show that $\operatorname{Div}_x \operatorname{Log}_x(y)$ can be expressed by $D(s)$.

Lemma 53. *Using previously defined $D(s)$, we have that*

$$\begin{aligned} \operatorname{Div}_x \operatorname{Log}_x(y) &= -1 - d(x,y) \left(\operatorname{tr} \left(\frac{d}{dt} D(t) \Big|_{t=d(x,y)} D(d(x,y))^{-1} \right) - \frac{1}{d(x,y)} \right) \\ &= -d(x,y) \operatorname{tr} \left(\frac{d}{dt} D(t) \Big|_{t=d(x,y)} D(d(x,y))^{-1} \right). \end{aligned}$$

Proof. Fix $s_0 \in (0, d(x,y))$. Choose vectors

$$e_1^{(s_0)}, \ldots, e_{d-1}^{(s_0)} \in v^\perp \subset T_y M$$

such that

$$J_{e_1^{(s_0)}}(s_0), \ldots, J_{e_{d-1}^{(s_0)}}(s_0)$$

form an orthonormal basis of a $d-1$ dimensional subspace of the tangent space at $T_{\gamma(s_0)} M$. Define

$$E_i(s_0) := J_{e_i^{(s_0)}}(s_0), \qquad i = 1, \ldots, d-1.$$

Since

$$J_{e_i^{(s_0)}}(0) = 0, \qquad D_t J_{e_i^{(s_0)}}(0) = e_i^{(s_0)}, \qquad \langle e_i^{(s_0)}, v \rangle = 0,$$

we know $J_{e_i^{(s_0)}}$ is a normal Jacobi field. Hence

$$\left\langle J_{e_i^{(s_0)}}(t), \gamma'(t) \right\rangle = 0 \qquad \text{for all } t.$$

In particular,

$$\langle E_i(s_0), \gamma'(s_0) \rangle = \left\langle J_{e_i^{(s_0)}}(s_0), \gamma'(s_0) \right\rangle = 0.$$

Now let

$$E_d(s_0) := \gamma'(s_0).$$

Then

$$E_1(s_0), \ldots, E_{d-1}(s_0), E_d(s_0)$$

is an orthonormal basis of $T_{\gamma(s_0)} M$.

We now compute

$$\nabla^2 d(\cdot, y)(\gamma(s_0))(E_i(s_0), E_i(s_0)).$$

Since

$$J_{e_i^{(s_0)}}(s_0) = E_i(s_0),$$

we obtain

$$\begin{aligned}\nabla^2 d(\cdot, y)(\gamma(s_0))(E_i(s_0), E_i(s_0)) &= \left\langle \nabla_{E_i(s_0)} \text{Grad } d(\cdot, y), E_i(s_0) \right\rangle \\ &= \left\langle \nabla_{J_{e_i^{(s_0)}}(s_0)} \text{Grad } d(\cdot, y), J_{e_i^{(s_0)}}(s_0) \right\rangle \\ &= \left\langle D_t J_{e_i^{(s_0)}}(s_0), J_{e_i^{(s_0)}}(s_0) \right\rangle.\end{aligned}$$

Next, by definition of $D(t)$,

$$P^y_{\gamma(t)} J_{e_i^{(s_0)}}(t) = D(t) e_i^{(s_0)},$$

hence

$$J_{e_i^{(s_0)}}(t) = P_y^{\gamma(t)} D(t) e_i^{(s_0)}.$$

Since $e_i^{(s_0)}$ is fixed with respect to t, differentiating along γ gives

$$D_t J_{e_i^{(s_0)}}(t) = P_y^{\gamma(t)} \frac{d}{dt} D(t) e_i^{(s_0)}.$$

Evaluating at $t = s_0$,

$$D_t J_{e_i^{(s_0)}}(s_0) = P_y^{\gamma(s_0)} \frac{d}{dt} D(t)\Big|_{t=s_0} e_i^{(s_0)}.$$

Therefore

$$\begin{aligned}\nabla^2 d(\cdot, y)(\gamma(s_0))(E_i(s_0), E_i(s_0)) &= \left\langle P_y^{\gamma(s_0)} \frac{d}{dt} D(t)\Big|_{t=s_0} e_i^{(s_0)}, P_y^{\gamma(s_0)} D(s_0) e_i^{(s_0)} \right\rangle \\ &= \left\langle \frac{d}{dt} D(t)\Big|_{t=s_0} e_i^{(s_0)}, D(s_0) e_i^{(s_0)} \right\rangle.\end{aligned}$$

Because

$$D(s_0) e_i^{(s_0)} = P^y_{\gamma(s_0)} J_{e_i^{(s_0)}}(s_0) = P^y_{\gamma(s_0)} E_i(s_0),$$

this becomes

$$\nabla^2 d(\cdot, y)(\gamma(s_0))(E_i(s_0), E_i(s_0)) = \left\langle \frac{d}{dt} D(t)\Big|_{t=s_0} D(s_0)^{-1} P^y_{\gamma(s_0)} E_i(s_0), P^y_{\gamma(s_0)} E_i(s_0) \right\rangle.$$

For the radial term, since Grad $d(\cdot, y)\big|_{\gamma(s_0)} = \gamma'(s_0)$, we have

$$\nabla^2 d(\cdot, y)(\gamma(s_0))(E_d(s_0), E_d(s_0)) = \left\langle \nabla_{\gamma'(s_0)} \text{Grad } d(\cdot, y), \gamma'(s_0) \right\rangle = \left\langle \nabla_{\gamma'(s_0)} \gamma'(s_0), \gamma'(s_0) \right\rangle = 0.$$

Therefore

$$\begin{aligned}\Delta d(\cdot, y)(\gamma(s_0)) &= \sum_{i=1}^{d} \nabla^2 d(\cdot, y)(\gamma(s_0))(E_i(s_0), E_i(s_0)) \\ &= \sum_{i=1}^{d-1} \left\langle \frac{d}{dt} D(t)\Big|_{t=s_0} D(s_0)^{-1} P^y_{\gamma(s_0)} E_i(s_0), P^y_{\gamma(s_0)} E_i(s_0) \right\rangle.\end{aligned}$$

On the other hand, since $P^y_{\gamma(s_0)} E_1(s_0), \ldots, P^y_{\gamma(s_0)} E_{d-1}(s_0), v$ is an orthonormal basis of $T_y M$, we have

$$\operatorname{tr}\left(\frac{d}{dt} D(t)\Big|_{t=s_0} D(s_0)^{-1} \right) = \sum_{i=1}^{d-1} \left\langle \frac{d}{dt} D(t)\Big|_{t=s_0} D(s_0)^{-1} P^y_{\gamma(s_0)} E_i(s_0), P^y_{\gamma(s_0)} E_i(s_0) \right\rangle$$

$$+ \left\langle \frac{d}{dt} D(t)\Big|_{t=s_0} D(s_0)^{-1} v, v \right\rangle.$$

Using the previously proved identity $D(t)v = tv$, we have

$$D(s_0)^{-1} v = \frac{1}{s_0} v, \qquad \frac{d}{dt} D(t)\Big|_{t=s_0} v = v.$$

Hence

$$\left\langle \frac{d}{dt} D(t)\Big|_{t=s_0} D(s_0)^{-1} v, v \right\rangle = \left\langle \frac{1}{s_0} \frac{d}{dt} D(t)\Big|_{t=s_0} v, v \right\rangle = \frac{1}{s_0}.$$

Therefore

$$\operatorname{tr}\left(\frac{d}{dt} D(t)\Big|_{t=s_0} D(s_0)^{-1} \right) = \Delta d(\cdot, y)(\gamma(s_0)) + \frac{1}{s_0}.$$

Equivalently,

$$\Delta d(\cdot, y)(\gamma(s_0)) = \operatorname{tr}\left(\frac{d}{dt} D(t)\Big|_{t=s_0} D(s_0)^{-1} \right) - \frac{1}{s_0}.$$

Now take

$$s_0 = d(x, y), \qquad x = \gamma(d(x, y)).$$

Recall $\operatorname{Div}_x \operatorname{Log}_x(y) = -\Delta_x \left(\frac{1}{2} d(x, y)^2 \right)$. By the product rule for divergence,

$$\begin{aligned} \Delta_x \left(\frac{1}{2} d(x, y)^2 \right) &= \operatorname{Div}_x \left(d(x, y) \operatorname{Grad}_x d(\cdot, y) \right) \\ &= \langle \operatorname{Grad}_x d(\cdot, y), \operatorname{Grad}_x d(\cdot, y) \rangle + d(x, y)\, \Delta_x d(\cdot, y). \end{aligned}$$

Since $x \neq y$,

$$\| \operatorname{Grad}_x d(\cdot, y) \| = 1.$$

Hence

$$\Delta_x \left(\frac{1}{2} d(x, y)^2 \right) = 1 + d(x, y) \Delta_x d(\cdot, y).$$

Therefore

$$\operatorname{Div}_x \operatorname{Log}_x(y) = -1 - d(x, y) \Delta_x d(\cdot, y).$$

Substituting

$$\Delta_x d(\cdot, y) = \operatorname{tr}\left(\frac{d}{dt} D(t)\Big|_{t=d(x,y)} D(d(x, y))^{-1} \right) - \frac{1}{d(x, y)},$$

we obtain

$$\begin{aligned} \operatorname{Div}_x \operatorname{Log}_x(y) &= -1 - d(x, y) \left(\operatorname{tr}\left(\frac{d}{dt} D(t)\Big|_{t=d(x,y)} D(d(x, y))^{-1} \right) - \frac{1}{d(x, y)} \right) \\ &= -d(x, y) \operatorname{tr}\left(\frac{d}{dt} D(t)\Big|_{t=d(x,y)} D(d(x, y))^{-1} \right). \end{aligned}$$

□

Now we state our bound.

Proposition 54.

$$\begin{aligned} \| \operatorname{Grad}_x \operatorname{Div}_x \operatorname{Log}_x(y) \| \leq d\, \frac{s'_{K_{\min}}(d(x, y))}{s_{K_{\min}}(d(x, y))} \left(d(x, y) \frac{s'_{K_{\min}}(d(x, y))}{s_{K_{\min}}(d(x, y))} - 1 \right) \\ + d(-K_{\min})\, d(x, y) + \frac{2 d L_R}{d(x, y)} s_{K_{\min}}(d(x, y))^2. \end{aligned}$$

Proof. Using Lemma 55 and Lemma 59,

$$\sup_{\substack{\xi\in T_xM,\|\xi\|=1,\\ \langle\xi,\gamma'(s_0)\rangle=0}} |\langle \text{Grad}_x \text{Div}_x \text{Log}_x(y),\xi\rangle| \le \frac{2dL_R}{d(x,y)^2}\int_0^{d(x,y)} s_{K_{\min}}(u)^2\,du, \le \frac{2dL_R}{d(x,y)} s_{K_{\min}}(d(x,y))^2$$

$$|\langle \text{Grad}_x \text{Div}_x \text{Log}_x(y),\gamma'(s_0)\rangle| \le d\,\frac{s'_{K_{\min}}(d(x,y))}{s_{K_{\min}}(d(x,y))}\left(d(x,y)\frac{s'_{K_{\min}}(d(x,y))}{s_{K_{\min}}(d(x,y))}-1\right)+d(-K_{\min})\,d(x,y).$$

So we can simply take

$$\begin{aligned}\|\,\text{Grad}_x \text{Div}_x \text{Log}_x(y)\| \le{}& d\,\frac{s'_{K_{\min}}(d(x,y))}{s_{K_{\min}}(d(x,y))}\left(d(x,y)\frac{s'_{K_{\min}}(d(x,y))}{s_{K_{\min}}(d(x,y))}-1\right)\\ &+d(-K_{\min})\,d(x,y)+\frac{2dL_R}{d(x,y)}s_{K_{\min}}(d(x,y))^2.\end{aligned}$$

□

E.6.1 Tangential Direction

We now bound the tangential derivative of $\text{Div}_x \text{Log}_x(y)$. We start with some notations. Let $\xi \in T_xM$ be such that $\|\xi\| = 1$ and $\langle\xi,\gamma'(s_0)\rangle = 0$. This ξ denotes the direction along which we want to evaluate the Riemannian gradient.

Since the map $e \mapsto J_e(s_0)$ is a linear bijection from $v^\perp \subset T_yM$ onto $\gamma'(s_0)^\perp \subset T_xM$, there exists a unique vector $\eta \in v^\perp$ such that $J_\eta(s_0) = \xi$. Such η can be viewed as, the initial tangent vector at y that "generate" the ξ direction at x. Now we construct a variation of geodesics. Let $w : (-\varepsilon,\varepsilon) \to T_yM$ be a smooth curve satisfying

$$\|w(\tau)\| = 1, \qquad w(0) = v, \qquad w'(0) = \eta.$$

Define a variation of geodesic as $\gamma_\tau(s) := \text{Exp}_y(s\,w(\tau))$. The requirement $\|w(\tau)\| = 1$ ensures $d(x(\tau),y) = s_0$ for every τ. For each τ and each $z \in T_yM$, let $J_{\tau,z}$ denote the Jacobi field along γ_τ satisfying

$$J_{\tau,z}(0) = 0, \qquad D_sJ_{\tau,z}(0) = z.$$

Then we know $\frac{\partial}{\partial\tau}\gamma_\tau(s)\big|_{\tau=0} = J_{\tau,\eta}$ (because $w'(0) = \eta$). Then denote $x(\tau) := \gamma_\tau(s_0)$, we have $x(0) = x$, and

$$\frac{d}{d\tau}x(\tau)\Big|_{\tau=0} = \frac{\partial}{\partial\tau}\gamma_\tau(s_0)\Big|_{\tau=0} = J_{\tau,\eta}(s_0) = \xi,$$

where the last equality is by definition.

Define

$$D_\tau(s)z := P^y_{\gamma_\tau(s)}J_{\tau,z}(s).$$

For some intuition, to compute the derivative at $x = \gamma(s_0)$ (where $s_0 = d(x,y)$) in a tangential direction, we keep the radius s_0 fixed and vary only the initial unit direction v of the radial geodesic from y. When $\tau = 0$, this reduces to the previously defined operator $D(s)z = P^y_{\gamma(s)}J_z(s)$. And as τ varies, $D_\tau(s)$ represent the variation of operator $D(s)$, with different initial conditions.

Lemma 55. *We have*

$$\sup_{\substack{\xi\in T_xM\\ \|\xi\|=1,\ \langle\xi,\gamma'(s_0)\rangle=0}} |\langle \textit{Grad}_x \textit{Div}_x \textit{Log}_x(y),\xi\rangle| \le \frac{2dL_R}{d(x,y)^2}\int_0^{d(x,y)} s_{K_{\min}}(u)^2\,du.$$

Proof.

By Lemma 53,

$$\text{Div}_x \text{Log}_x(y) = -d(x,y)\,\text{tr}\left(\frac{d}{dt}D(t)\Big|_{t=d(x,y)}D(d(x,y))^{-1}\right).$$

Since $d(x(\tau), y) = s_0$ for every τ, the above naturally extends to operators $D_\tau(s)$, i.e., we have

$$\operatorname{Div}_{x(\tau)} \operatorname{Log}_{x(\tau)}(y) = -s_0 \operatorname{tr}\left(\frac{d}{ds} D_\tau(s)\Big|_{s=s_0} D_\tau(s_0)^{-1}\right).$$

Therefore (by definition of Riemannian gradient)

$$\begin{aligned}\langle \operatorname{Grad}_x \operatorname{Div}_x \operatorname{Log}_x(y), \xi\rangle &= \frac{d}{d\tau} \operatorname{Div}_{x(\tau)} \operatorname{Log}_{x(\tau)}(y)\Big|_{\tau=0} \\ &= -s_0 \frac{d}{d\tau} \operatorname{tr}\left(\frac{d}{ds} D_\tau(s)\Big|_{s=s_0} D_\tau(s_0)^{-1}\right)\Big|_{\tau=0}.\end{aligned}$$

Using $|\operatorname{tr}(A)| \le d\,\|A\|$ (for any endomorphism A of the d-dimensional space T_yM) and Lemma 56, we have

$$\begin{aligned}|\langle \operatorname{Grad}_x \operatorname{Div}_x \operatorname{Log}_x(y), \xi\rangle| &= \left|-s_0 \operatorname{tr}\left(\frac{d}{d\tau}\left(\frac{d}{ds} D_\tau(s)\Big|_{s=s_0} D_\tau(s_0)^{-1}\right)\Big|_{\tau=0}\right)\right| \\ &\le s_0\, d \left\|\frac{d}{d\tau}\left(\frac{d}{ds} D_\tau(s)\Big|_{s=s_0} D_\tau(s_0)^{-1}\right)\Big|_{\tau=0}\right\| \\ &= s_0\, d \Bigg\| - D(s_0)^{-1}\left[\int_0^{s_0} D(u)\big(R(D(u)D(s_0)^{-1}(\cdot),\eta)v\big)\,du\right] \\ &\qquad\qquad - D(s_0)^{-1}\left[\int_0^{s_0} D(u)\big(R(D(u)D(s_0)^{-1}(\cdot),v)\eta\big)\,du\right]\Bigg\| \\ &\le s_0\, d \left\|D(s_0)^{-1}\left[\int_0^{s_0} D(u)\big(R(D(u)D(s_0)^{-1}(\cdot),\eta)v\big)\,du\right]\right\| \\ &\quad + s_0\, d \left\|D(s_0)^{-1}\left[\int_0^{s_0} D(u)\big(R(D(u)D(s_0)^{-1}(\cdot),v)\eta\big)\,du\right]\right\| \\ &\le s_0\, d\, \|D(s_0)^{-1}\|^2 \left\|\int_0^{s_0} D(u)\big(R(D(u)(\cdot),\eta)v\big)\,du\right\| \\ &\quad + s_0\, d\, \|D(s_0)^{-1}\|^2 \left\|\int_0^{s_0} D(u)\big(R(D(u)(\cdot),v)\eta\big)\,du\right\|.\end{aligned}$$

We now estimate the two operator norms. For every unit vector $z \in T_yM$, the curvature bound gives

$$\|R(D(u)z,\eta)v\| \le L_R\,\|D(u)z\|\,\|\eta\|\,\|v\| = L_R\,\|D(u)z\|\,\|\eta\|,$$

and similarly

$$\|R(D(u)z,v)\eta\| \le L_R\,\|D(u)z\|\,\|v\|\,\|\eta\| = L_R\,\|D(u)z\|\,\|\eta\|.$$

Hence

$$\begin{aligned}\left\|\int_0^{s_0} D(u)\big(R(D(u)\cdot,\eta)v\big)\,du\right\| &\le L_R\,\|\eta\| \int_0^{s_0} \|D(u)\|^2\,du, \\ \left\|\int_0^{s_0} D(u)\big(R(D(u)\cdot,v)\eta\big)\,du\right\| &\le L_R\,\|\eta\| \int_0^{s_0} \|D(u)\|^2\,du.\end{aligned}$$

Therefore

$$|\langle \operatorname{Grad}_x \operatorname{Div}_x \operatorname{Log}_x(y), \xi\rangle| \le 2dL_R\, s_0\, \|D(s_0)^{-1}\|^2\,\|\eta\| \int_0^{s_0} \|D(u)\|^2\,du.$$

By Rauch comparison theorem (Jacobi field comparison), since γ is a unit-speed geodesic and J is a normal Jacobi field along γ with $J(0) = 0$, we get

$$\|J(u)\| \le s_{K_{\min}}(u)\|D_tJ(0)\|,$$

where $s_{K_{\min}}(t) = \frac{1}{\sqrt{-K_{\min}}} \sinh t\sqrt{-K_{\min}}$ is the model function. Also, if sectional curvature is upper bounded by 0, we have $\|J(u)\| \ge u\|D_tJ(0)\|$.

We now apply these two conclusions to the Jacobi fields defining $D(u)$. Recall that the Jacobi field J_z is normal, and $D(u)z = P^y_{\gamma(u)} J_z(u)$ by definition. Since parallel transport is isometry, we obtain $\|D(u)z\| = \|J_z(u)\|$.

Thus, Rauch comparison theorem gives

$$\|D(u)z\| = \|J_z(u)\| \le s_{K_{\min}}(u).$$

Taking the supremum over all unit vectors $z \in v^\perp$ gives

$$\|D(u)\| \le s_{K_{\min}}(u).$$

Next, applying Rauch comparison under $\mathrm{Sec} \le 0$, we obtain

$$\|D(u)z\| = \|J_z(u)\| \ge u \qquad \text{for every } z \in v^\perp \text{ with } \|z\| = 1.$$

Evaluating at $u = s_0$, we get

$$\|D(s_0)^{-1}\| \le \frac{1}{s_0}.$$

Finally, since

$$\xi = J_\eta(s_0) = P_y^{\gamma(s_0)} D(s_0)\eta, \qquad \|\xi\| = 1,$$

we have $\|D(s_0)\eta\| = 1$. Since $\|D(s_0)\eta\| = \|J_\eta(s_0)\| \ge s_0 \|D_s J_\eta(0)\| = s_0\|\eta\|$ we conclude that

$$\|\eta\| \le \frac{1}{s_0}\|D(s_0)\eta\| = \frac{1}{s_0}.$$

Substituting these estimates into the previous bound yields

$$|\langle \mathrm{Grad}_x \mathrm{Div}_x \mathrm{Log}_x(y), \xi\rangle| \le \frac{2dL_R}{s_0^2} \int_0^{s_0} s_{K_{\min}}(u)^2\, du.$$

Since ξ was an arbitrary unit vector in $\gamma'(s_0)^\perp$, we conclude that

$$\sup_{\substack{\xi \in T_x M \\ \|\xi\|=1,\ \langle \xi, \gamma'(s_0)\rangle = 0}} |\langle \mathrm{Grad}_x \mathrm{Div}_x \mathrm{Log}_x(y), \xi\rangle| \le \frac{2dL_R}{d(x,y)^2} \int_0^{d(x,y)} s_{K_{\min}}(u)^2\, du.$$

□

We next compute the derivative with respect to τ.

Lemma 56. *We have that*

$$\begin{aligned} \frac{d}{d\tau}\left(\frac{d}{ds} D_\tau(s)\Big|_{s=s_0} D_\tau(s_0)^{-1}\right)\Big|_{\tau=0} z = &-D(s_0)^{-1}\left[\int_0^{s_0} D(u)\big(R(D(u)D(s_0)^{-1}z, \eta)v\big)\, du\right] \\ &- D(s_0)^{-1}\left[\int_0^{s_0} D(u)\big(R(D(u)D(s_0)^{-1}z, v)\eta\big)\, du\right]. \end{aligned} \tag{33}$$

Proof.

Fix $z \in T_y M$. We consider $\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}$. For each τ, recall that $D_\tau(s)z := P^y_{\gamma_\tau(s)} J_{\tau,z}(s)$, and $\gamma_\tau(s) = \mathrm{Exp}_y(s\, w(\tau))$. So we see the curve along which the Jacobi field is associated to, has a time derivative as $w(\tau)$ (up to parallel transport). Thus equation (32) becomes

$$\frac{d^2}{ds^2}\big(D_\tau(s)z\big) + R(D_\tau(s)z, w(\tau))w(\tau) = 0.$$

We now differentiate this identity with respect to τ at $\tau = 0$.

For the first term, differentiation with respect to τ commutes with differentiation with respect to s, so

$$\frac{d}{d\tau}\left(\frac{d^2}{ds^2}\big(D_\tau(s)z\big)\right)\Big|_{\tau=0} = \frac{d^2}{ds^2}\left(\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}\right).$$

For the curvature term, note that after parallel transport back to T_yM, the curvature tensor is evaluated at the fixed point y, and hence

$$(u_1, u_2, u_3) \mapsto R(u_1, u_2)u_3$$

is a trilinear map on the fixed vector space T_yM. Therefore, using the product rule and recall

$$D_0(s) = D(s), \qquad w(0) = v, \qquad w'(0) = \eta,$$

we obtain

$$\begin{aligned}\frac{d}{d\tau}\Big(R(D_\tau(s)z, w(\tau))w(\tau)\Big)\Big|_{\tau=0} = R\left(\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}, v\right)v \\ + R(D(s)z, \eta)v \\ + R(D(s)z, v)\eta.\end{aligned}$$

Hence we obtain

$$\frac{d^2}{ds^2}\left(\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}\right) + R\left(\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}, v\right)v + R(D(s)z, \eta)v + R(D(s)z, v)\eta = 0.$$

We next record the initial conditions. Since $D_\tau(0)z = 0, \forall\tau$, differentiating with respect to τ gives

$$\frac{d}{d\tau}\big(D_\tau(0)z\big)\Big|_{\tau=0} = 0.$$

Also, for every τ,

$$\frac{d}{ds}\big(D_\tau(s)z\big)\Big|_{s=0} = \frac{d}{ds}\big(P^y_{\gamma_\tau(s)}J_{\tau,z}(s)\big)\Big|_{s=0} = D_sJ_{\tau,z}(0) = z.$$

Since the right-hand side z does not depend on τ, differentiating with respect to τ gives

$$\frac{d}{ds}\left(\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}\right)\Big|_{s=0} = 0.$$

Next, we differentiate

$$\frac{d}{ds}D_\tau(s)\Big|_{s=s_0}D_\tau(s_0)^{-1}.$$

Recall the inverse derivative formula:

$$\begin{aligned}\frac{d}{d\tau}\left(D_\tau(s_0)^{-1}\right)\Big|_{\tau=0} &= -D_0(s_0)^{-1}\left(\frac{d}{d\tau}D_\tau(s_0)\Big|_{\tau=0}\right)D_0(s_0)^{-1} \\ &= -D(s_0)^{-1}\left(\frac{d}{d\tau}D_\tau(s_0)\Big|_{\tau=0}\right)D(s_0)^{-1}.\end{aligned}$$

Fix any $z \in T_yM$. We compute

$$\begin{aligned}&\frac{d}{d\tau}\left[\left(\frac{d}{ds}D_\tau(s)\Big|_{s=s_0}D_\tau(s_0)^{-1}\right)z\right]\Big|_{\tau=0} \\ &= \frac{d}{d\tau}\left[\frac{d}{ds}D_\tau(s)\Big|_{s=s_0}\big(D_\tau(s_0)^{-1}z\big)\right]\Big|_{\tau=0} \\ &= \left(\frac{d}{d\tau}\left(\frac{d}{ds}D_\tau(s)\Big|_{s=s_0}\right)\Big|_{\tau=0}\right)\big(D(s_0)^{-1}z\big) + \frac{d}{ds}D(s)\Big|_{s=s_0}\left(\frac{d}{d\tau}\left(D_\tau(s_0)^{-1}z\right)\Big|_{\tau=0}\right) \\ &= \frac{d}{ds}\left(\frac{d}{d\tau}D_\tau(s)\Big|_{\tau=0}\right)\Big|_{s=s_0}\big(D(s_0)^{-1}z\big) - \frac{d}{ds}D(s)\Big|_{s=s_0}\left[D(s_0)^{-1}\left(\frac{d}{d\tau}D_\tau(s_0)\Big|_{\tau=0}\right)D(s_0)^{-1}z\right].\end{aligned}$$

From Lemma 58 we have

$$D(s_0)^{-1}\frac{d}{ds}D(s)\Big|_{s=s_0} = \frac{d}{ds}D(s)\Big|_{s=s_0} D(s_0)^{-1},$$

so we get

$$\begin{aligned}
&\frac{d}{d\tau}\left[\left(\frac{d}{ds}D_\tau(s)\Big|_{s=s_0} D_\tau(s_0)^{-1}\right) z\right]\Big|_{\tau=0} \\
&= D(s_0)^{-1}\left[D(s_0)\frac{d}{ds}\left(\frac{d}{d\tau}D_\tau(s)\Big|_{\tau=0}\right)\Big|_{s=s_0}\big(D(s_0)^{-1}z\big) - \frac{d}{ds}D(s)\Big|_{s=s_0}\left(\left(\frac{d}{d\tau}D_\tau(s_0)\Big|_{\tau=0}\right)D(s_0)^{-1}z\right)\right] \\
&= D(s_0)^{-1}\left[D(s_0)\frac{d}{ds}\left(\frac{d}{d\tau}\big(D_\tau(s)\big(D(s_0)^{-1}z\big)\big)\Big|_{\tau=0}\right)\Big|_{s=s_0} - \frac{d}{ds}D(s)\Big|_{s=s_0}\left(\frac{d}{d\tau}\big(D_\tau(s_0)\big(D(s_0)^{-1}z\big)\big)\Big|_{\tau=0}\right)\right].
\end{aligned}$$

We apply Lemma 57 with $\tilde{z} = D(s_0)^{-1}z$. Substituting the integrated identity above, we obtain

$$\begin{aligned}
\frac{d}{d\tau}\left(\frac{d}{ds}D_\tau(s)\Big|_{s=s_0} D_\tau(s_0)^{-1}\right)\Big|_{\tau=0} z = &-D(s_0)^{-1}\left[\int_0^{s_0} D(u)\big(R(D(u)D(s_0)^{-1}z,\eta)v\big)\,du\right] \\
&- D(s_0)^{-1}\left[\int_0^{s_0} D(u)\big(R(D(u)D(s_0)^{-1}z,v)\eta\big)\,du\right].
\end{aligned}$$

□

Lemma 57. *Fix any $\tilde{z} \in T_yM$. We have*

$$\begin{aligned}
&D(s_0)\frac{d}{ds}\left(\frac{d}{d\tau}\big(D_\tau(s)\tilde{z}\big)\Big|_{\tau=0}\right)\Big|_{s=s_0} - \frac{d}{ds}D(s)\Big|_{s=s_0}\left(\frac{d}{d\tau}\big(D_\tau(s_0)\tilde{z}\big)\Big|_{\tau=0}\right) \\
&= -\int_0^{s_0} D(u)\big(R(D(u)\tilde{z},\eta)v\big)\,du - \int_0^{s_0} D(u)\big(R(D(u)\tilde{z},v)\eta\big)\,du.
\end{aligned}$$

Proof.

For ease of notation, we write $\tilde{z}$ as z throughout the proof. Define

$$A_z(s) := D(s)\frac{d}{ds}\left(\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}\right) - \frac{d}{ds}D(s)\left(\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}\right).$$

We differentiate this quantity with respect to s. Using the product rule, we obtain

$$\begin{aligned}
\frac{d}{ds}A_z(s) &= \frac{d}{ds}D(s)\,\frac{d}{ds}\left(\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}\right) + D(s)\,\frac{d^2}{ds^2}\left(\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}\right) \\
&\quad - \frac{d^2}{ds^2}D(s)\left(\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}\right) - \frac{d}{ds}D(s)\,\frac{d}{ds}\left(\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}\right) \\
&= D(s)\,\frac{d^2}{ds^2}\left(\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}\right) - \frac{d^2}{ds^2}D(s)\left(\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}\right).
\end{aligned}$$

For the first term, recall that

$$\frac{d^2}{ds^2}\left(\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}\right) = -R\left(\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}, v\right)v - R(D(s)z,\eta)v - R(D(s)z,v)\eta.$$

For the second term, since $D(s)$ satisfies $\frac{d^2}{ds^2}\big(D(s)u\big) + R(D(s)u,v)v = 0$ for all $u \in T_yM$, we take $u = \frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}$ and obtain

$$\frac{d^2}{ds^2}D(s)\left(\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}\right) = -R\left(D(s)\left(\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}\right), v\right)v.$$

Substitute back, we obtain

$$\begin{aligned}\frac{d}{ds}A_z(s) = &-D(s)\left(R\left(\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}, v\right)v\right)\\ &- D(s)\big(R(D(s)z,\eta)v\big) - D(s)\big(R(D(s)z,v)\eta\big)\\ &+ R\left(D(s)\left(\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}\right), v\right)v.\end{aligned}$$

By Lemma 58, $D(s)R(\cdot,v)v = R(D(s)\cdot,v)v$. This implies

$$D(s)\left(R\left(\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}, v\right)v\right) = R\left(D(s)\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}, v\right)v,$$

so the first and the last terms cancel, and we obtain

$$\frac{d}{ds}A_z(s) = -D(s)\big(R(D(s)z,\eta)v\big) - D(s)\big(R(D(s)z,v)\eta\big).$$

By definition of $A_z(s)$,

$$A_z(0) = D(0)\,\frac{d}{ds}\left(\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}\right)\Big|_{s=0} - \frac{d}{ds}D(0)\,\left(\frac{d}{d\tau}\big(D_\tau(0)z\big)\Big|_{\tau=0}\right).$$

Using

$$D(0) = 0, \qquad \frac{d}{d\tau}\big(D_\tau(0)z\big)\Big|_{\tau=0} = 0,$$

we get

$$A_z(0) = 0.$$

Thus, integrating the identity for $\frac{d}{ds}A_z(s)$ from 0 to s_0 gives

$$A_z(s_0) = A_z(s_0) - A_z(0) = -\int_0^{s_0} D(u)\big(R(D(u)z,\eta)v\big)\,du - \int_0^{s_0} D(u)\big(R(D(u)z,v)\eta\big)\,du.$$

Equivalently,

$$\begin{aligned}&D(s_0)\frac{d}{ds}\left(\frac{d}{d\tau}\big(D_\tau(s)z\big)\Big|_{\tau=0}\right)\Big|_{s=s_0} - \frac{d}{ds}D(s)\Big|_{s=s_0}\left(\frac{d}{d\tau}\big(D_\tau(s_0)z\big)\Big|_{\tau=0}\right)\\ &= -\int_0^{s_0} D(u)\big(R(D(u)z,\eta)v\big)\,du - \int_0^{s_0} D(u)\big(R(D(u)z,v)\eta\big)\,du.\end{aligned}$$

□

Lemma 58.

$$D(s_0)^{-1}\frac{d}{ds}D(s)\Big|_{s=s_0} = \frac{d}{ds}D(s)\Big|_{s=s_0}D(s_0)^{-1}.$$

Also, $D(s)R(\cdot,v)v = R(D(s)\cdot,v)v$.

Proof. Notice that

$$\mathrm{Rm}\,(w,v,v,u) = -\,\mathrm{Rm}\,(v,w,v,u) = \mathrm{Rm}\,(v,w,u,v) = \mathrm{Rm}\,(u,v,v,w),$$

so we have $\langle R(w,v)v,u\rangle = \langle R(u,v)v,w\rangle = \langle w,R(u,v)v\rangle$. Thus we can view $R(\cdot,v)v$ as a self-adjoint linear operator. Consequently, there exists $b_1,...,b_d$ s.t. $R(b_i,v)v = \lambda_i b_i, \forall i$, for some $\lambda_1,...,\lambda_d$. Now consider $D(s)b_i$. The Jacobi equation is

$$\frac{d^2}{ds^2}(D(s)b_i) + R(D(s)b_i,v)v = 0.$$

Notice that its initial values are $D(0) = 0$ and $D'(0) = \mathrm{Id}$. Consider the scalar-valued ODE

$$f_i''(s) + \lambda_i f_i(s) = 0, \quad f_i(0) = 0, \quad f_i'(0) = 1.$$

Its solution $f_i(s)$ satisfies $R(f_i(s)b_i, v)v = f_i(s)R(b_i, v)v = f_i(s)\lambda_i b_i$ and $f_i''(s)b_i = \frac{d^2}{ds^2}(f_i(s)b_i)$. So we know

$$\frac{d^2}{ds^2}(f_i(s)b_i) + R(f_i(s)b_i, v)v = (f_i''(s) + \lambda_i f_i(s))b_i = 0.$$

Hence we know $D(s)b_i = f_i(s)b_i$ (by uniqueness of ODE solution).

Under the basis b_i, $D(s)$ can be represented by a scalar valued function $f_i(s)$, so it's diagonal. Therefore,

$$D(s_0)^{-1}\frac{d}{ds}D(s)\Big|_{s=s_0} = \frac{d}{ds}D(s)\Big|_{s=s_0} D(s_0)^{-1}.$$

We also get,

$$D(s)R(b_i, v)v = R(D(s)b_i, v)v, \forall i,$$

so consequently we also have $D(s)R(\cdot, v)v = R(D(s)\cdot, v)v$. □

E.6.2 Radial Direction

For the radial derivative, fix

$$x = \gamma(s_0), \qquad s_0 = d(x, y).$$

Define

$$f(s) := \mathrm{Div}_{\gamma(s)} \mathrm{Log}_{\gamma(s)}(y).$$

Lemma 59.

$$|\langle \mathit{Grad}_x \mathit{Div}_x \mathit{Log}_x(y), \gamma'(s_0)\rangle| \le d\,\frac{s'_{K_{\min}}(d(x,y))}{s_{K_{\min}}(d(x,y))}\left(d(x,y)\frac{s'_{K_{\min}}(d(x,y))}{s_{K_{\min}}(d(x,y))} - 1\right) + d(-K_{\min})\,d(x,y).$$

Proof.

By Lemma 53,

$$f(s) = -s\ \mathrm{tr}\left(\frac{d}{ds}D(s)D(s)^{-1}\right).$$

Since $\gamma'(s_0)$ is the velocity of the curve $s \mapsto \gamma(s)$ at $s = s_0$, we have

$$\langle \mathrm{Grad}_x \mathrm{Div}_x \mathrm{Log}_x(y), \gamma'(s_0)\rangle = \frac{d}{ds}f(s)\Big|_{s=s_0}.$$

Therefore it suffices to compute $\frac{d}{ds}f(s)$.

We differentiate $f(s)$ with respect to s:

$$\begin{aligned}\frac{d}{ds}f(s) &= -\mathrm{tr}\left(\frac{d}{ds}D(s)D(s)^{-1}\right) - s\,\frac{d}{ds}\mathrm{tr}\left(\frac{d}{ds}D(s)D(s)^{-1}\right)\\ &= -\mathrm{tr}\left(\frac{d}{ds}D(s)D(s)^{-1}\right) - s\ \mathrm{tr}\left(\frac{d}{ds}\left(\frac{d}{ds}D(s)D(s)^{-1}\right)\right).\end{aligned}$$

From the established Riccati equation (Lemma 50),

$$\frac{d}{ds}\left(\frac{d}{ds}D(s)D(s)^{-1}\right) + \left(\frac{d}{ds}D(s)D(s)^{-1}\right)\circ\left(\frac{d}{ds}D(s)D(s)^{-1}\right) + \big(z \mapsto R(z, v)v\big) = 0,$$

we obtain

$$\frac{d}{ds}\left(\frac{d}{ds}D(s)D(s)^{-1}\right) = -\left(\frac{d}{ds}D(s)D(s)^{-1}\right)\circ\left(\frac{d}{ds}D(s)D(s)^{-1}\right) - \big(z \mapsto R(z, v)v\big).$$

Substituting this into the previous identity gives

$$\begin{aligned}\frac{d}{ds}f(s) = &-\operatorname{tr}\left(\frac{d}{ds}D(s)D(s)^{-1}\right)\\ &+ s\ \operatorname{tr}\left(\left(\frac{d}{ds}D(s)D(s)^{-1}\right)\circ\left(\frac{d}{ds}D(s)D(s)^{-1}\right)\right)\\ &+ s\ \operatorname{tr}\big(z\mapsto R(z,v)v\big).\end{aligned}$$

We now diagonalize the operator $\frac{d}{ds}D(s)D(s)^{-1}$. First, from the identity

$$D(s)v = sv,$$

we differentiate with respect to s and obtain

$$\frac{d}{ds}D(s)v = v.$$

Hence

$$\left(\frac{d}{ds}D(s)D(s)^{-1}\right)v = \frac{1}{s}v.$$

As in Lemma 52, $\frac{d}{ds}D(s)D(s)^{-1}$ can be viewed as the Hessian of the distance function on the subspace $v^{\perp}$ (up to parallel transport), the restriction of

$$\frac{d}{ds}D(s)D(s)^{-1}$$

to $v^{\perp}$ is self-adjoint. Therefore there exists an orthonormal eigenvectors $e_1(s),\dots,e_{d-1}(s)$ in $v^{\perp}$ and eigenvalues $\lambda_1(s),\dots,\lambda_{d-1}(s)$ such that

$$\left(\frac{d}{ds}D(s)D(s)^{-1}\right)e_i(s) = \lambda_i(s)e_i(s), \qquad i = 1,\dots,d-1.$$

Thus, with respect to the orthonormal basis

$$e_1(s),\dots,e_{d-1}(s),v$$

of T_yM, the operator $\frac{d}{ds}D(s)D(s)^{-1}$ has eigenvalues

$$\lambda_1(s),\dots,\lambda_{d-1}(s),\frac{1}{s}.$$

Hence

$$\operatorname{tr}\left(\frac{d}{ds}D(s)D(s)^{-1}\right) = \sum_{i=1}^{d-1}\lambda_i(s) + \frac{1}{s},$$

and

$$\operatorname{tr}\left(\left(\frac{d}{ds}D(s)D(s)^{-1}\right)\circ\left(\frac{d}{ds}D(s)D(s)^{-1}\right)\right) = \sum_{i=1}^{d-1}\lambda_i(s)^2 + \frac{1}{s^2}.$$

Substituting these two identities into the formula for $\frac{d}{ds}f(s)$ yields

$$\begin{aligned}\frac{d}{ds}f(s) &= -\left(\sum_{i=1}^{d-1}\lambda_i(s) + \frac{1}{s}\right) + s\left(\sum_{i=1}^{d-1}\lambda_i(s)^2 + \frac{1}{s^2}\right) + s\ \operatorname{tr}\big(z\mapsto R(z,v)v\big)\\ &= \sum_{i=1}^{d-1}\big(-\lambda_i(s) + s\lambda_i(s)^2\big) + s\ \operatorname{tr}\big(z\mapsto R(z,v)v\big).\end{aligned}$$

Equivalently,

$$\frac{d}{ds}f(s) = \sum_{i=1}^{d-1} \lambda_i(s)\big(s\lambda_i(s) - 1\big) + s\ \mathrm{tr}\big(z \mapsto R(z,v)v\big).$$

We now bound the first term by comparison theory. By Hessian comparison under

$$K_{\min} \le \mathrm{Sec} \le 0,$$

the eigenvalues $\lambda_i(s)$ satisfy

$$\frac{1}{s} \le \lambda_i(s) \le \frac{s'_{K_{\min}}(s)}{s_{K_{\min}}(s)}, \qquad i = 1, \dots, d-1.$$

Therefore

$$s\lambda_i(s) - 1 \ge 0,$$

and also

$$\lambda_i(s)\big(s\lambda_i(s) - 1\big) \le \frac{s'_{K_{\min}}(s)}{s_{K_{\min}}(s)} \left(s\frac{s'_{K_{\min}}(s)}{s_{K_{\min}}(s)} - 1 \right).$$

Summing over $i = 1, \dots, d-1$, we obtain

$$\begin{aligned}\sum_{i=1}^{d-1} \lambda_i(s)\big(s\lambda_i(s) - 1\big) &\le (d-1)\frac{s'_{K_{\min}}(s)}{s_{K_{\min}}(s)} \left(s\frac{s'_{K_{\min}}(s)}{s_{K_{\min}}(s)} - 1 \right) \\ &\le d\,\frac{s'_{K_{\min}}(s)}{s_{K_{\min}}(s)} \left(s\frac{s'_{K_{\min}}(s)}{s_{K_{\min}}(s)} - 1 \right).\end{aligned}$$

We next bound the curvature trace term. Since

$$\mathrm{tr}\big(z \mapsto R(z,v)v\big) = \sum_{i=1}^{d-1} \langle R(e_i(s), v)v, e_i(s)\rangle\,,$$

and each quantity

$$\langle R(e_i(s), v)v, e_i(s)\rangle$$

is the sectional curvature of the plane spanned by $e_i(s)$ and v, we have

$$K_{\min} \le \langle R(e_i(s), v)v, e_i(s)\rangle \le 0.$$

Hence

$$\begin{aligned}\big|\mathrm{tr}\big(z \mapsto R(z,v)v\big)\big| &\le (d-1)(-K_{\min}) \\ &\le d(-K_{\min}).\end{aligned}$$

Therefore

$$\big|s\ \mathrm{tr}\big(z \mapsto R(z,v)v\big)\big| \le d(-K_{\min})\, s.$$

Combining the two bounds, we conclude that

$$\left|\frac{d}{ds}f(s)\right| \le d\,\frac{s'_{K_{\min}}(s)}{s_{K_{\min}}(s)} \left(s\frac{s'_{K_{\min}}(s)}{s_{K_{\min}}(s)} - 1 \right) + d(-K_{\min})\, s.$$

Evaluating at $s = s_0 = d(x,y)$ gives

$$\begin{aligned}|\langle \mathrm{Grad}_x\, \mathrm{Div}_x\, \mathrm{Log}_x(y), \gamma'(s_0)\rangle| \le\ & d\,\frac{s'_{K_{\min}}(d(x,y))}{s_{K_{\min}}(d(x,y))} \left(d(x,y)\frac{s'_{K_{\min}}(d(x,y))}{s_{K_{\min}}(d(x,y))} - 1 \right) \\ &+ d(-K_{\min})\, d(x,y).\end{aligned}$$

□

E.7 Auxiliary results

Given a matrix $A(x)$, the directional derivative in v can be computed through

$$D \log\det(A(x))[v] = \operatorname{tr}(A^{-1} DA(x)[v]).$$

Hence

$$\|\nabla \log\det(A(x))\|_{\mathrm{op}} = \sup_{\|v\|=1} |\operatorname{tr}(A^{-1} DA(x)[v])| \le d\|A^{-1}\| \times \|DA(x)\|_{\mathrm{op}}.$$

Furthermore, using the product rule, and noting that $DA^{-1}(x) = -A^{-1}(x)DA(x)A^{-1}(x)$, we obtain

$$\begin{aligned}
&D^2 \log\det(A(x))[v,v] \\
&= \operatorname{tr}(DA^{-1}(x)[v] DA(x)[v]) + \operatorname{tr}(A^{-1}(x) D^2 A(x)[v,v]) \\
&= -\operatorname{tr}(A^{-1}(x) DA(x)[v] A^{-1}(x) DA(x)[v]) + \operatorname{tr}(A^{-1}(x) D^2 A(x)[v,v]).
\end{aligned}$$

Hence, we have

$$\|\nabla^2 \log\det(A(x))\|_{op} \le d\|A^{-1}(x)\|^2 \times \|\nabla A(x)\|^2 + d\|A^{-1}(x)\| \times \|\nabla^2 A(x)\|.$$

This is summarized as the following result, and is used in Lemma 42.

Lemma 60 (Derivatives of $\log\det$)**.** *Let $A(x)$ be a smooth family of invertible linear maps on a d-dimensional inner product space. Then for any unit vector v,*

$$D \log\det(A(x))[v] = \operatorname{tr}\left(A(x)^{-1} DA(x)[v]\right).$$

Consequently,

$$\|\nabla \log\det(A(x))\| = \sup_{\|v\|=1} \left|\operatorname{tr}\left(A(x)^{-1} DA(x)[v]\right)\right| \le d\,\|A(x)^{-1}\|\,\|\nabla A(x)\|, \tag{34}$$

$$\|\nabla^2 \log\det(A(x))\|_{\mathrm{op}} \le d\,\|A(x)^{-1}\|^2\,\|\nabla A(x)\|^2 + d\,\|A(x)^{-1}\|\,\|\nabla^2 A(x)\|. \tag{35}$$

F LLM usage statement

LLM-based tools were used during manuscript preparation for limited brainstorming, drafting, editing, and formatting assistance, including preliminary drafts of some proof arguments. The LLM-generated material was not used without author review. All proof arguments were independently checked, corrected, and finalized by the authors, who take full responsibility for the correctness, originality, and presentation of the paper.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We established the total variation error bound and iteration complexity. The statements are in Section 3 and proofs are provided in the Appendix.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: For our theoretical analysis, on compact manifolds, we provided explicit rate for hyperspheres. We discussed the main difficulty faced by the general compact manifold case (cut locus problem) in section 4.2.1.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: The assumptions are in Section 3 and proofs are provided in the Appendix. We provided a proof sketch in Section 4.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [N/A]

   Justification: the paper does not include experiments.

   Guidelines:

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

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

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

Answer: [N/A]

Justification: paper does not include experiments requiring code

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [N/A]

   Justification: the paper does not include experiments.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [N/A]

   Justification: the paper does not include experiments.

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [N/A]

   Justification: the paper does not include experiments.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

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

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: Our work is theoretical.

    Guidelines:

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

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

11. **Safeguards**

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

    Answer: [N/A]

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

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

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

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

13. **New assets**

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

    Answer: [N/A]

    Justification: the paper does not release new assets.

    Guidelines:

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

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

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

    Answer: [N/A]

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

    Guidelines:

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

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

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

    Answer: [N/A]

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

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [Yes]

Justification: We declared LLM usage in Appendix F. The LLM-generated material was independently checked, corrected, and finalized by the authors, who take full responsibility for the correctness, originality, and presentation of the paper.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.