Riemannian Score-Based Generative Modeling

Anonymous authors
Paper under double-blind review

Abstract

Score-based generative models (SGMs) are a novel class of generative models demonstrating remarkable empirical performance. One uses a diffusion to add progressively Gaussian noise to the data, while the generative model is a "denoising" process obtained by approximating the time-reversal of this "noising" diffusion. However, current SGMs make the underlying assumption that the data is supported on a Euclidean manifold with flat geometry. This prevents the use of these models for applications in robotics, geoscience or protein modeling which rely on distributions defined on Riemannian manifolds. To overcome this issue, we introduce *Riemannian Score-based Generative Models* (RSGMs) which extend current SGMs to the setting of compact Riemannian manifolds. We also show how RSGMs can be accelerated by solving a Schrödinger bridge problem on manifolds. We illustrate our approach with earth and climate science data.

Keywords— Diffusion processes, Generative modeling, Riemannian manifold, Score-based generative models, Schrödinger bridge

1 Introduction

Score-based Generative Modeling (SGM) is a recently developed approach to generative modeling exhibiting state-of-the-art performances on various tasks including image and audio synthesis D(Song and Ermon, 2019; Song et al., 2021; Ho et al., 2020; Nichol and Dhariwal, 2021; Dhariwal and Nichol, 2021). These models proceed as follows. We add noise to the data progressively using a diffusion process targeting a reference Gaussian distribution. The corresponding time-reversal process is also a diffusion whose drift depends on the logarithmic gradients of the perturbed data distributions, i.e. the scores. The generative model is obtained by approximating this time-reversal denoising diffusion by initializing it at the reference Gaussian distribution and using neural networks estimates of the scores obtained using score matching Hyvärinen and Dayan (2005); Vincent (2011). It can be shown rigorously that the obtained final samples are approximately distributed according to the data distribution (De Bortoli et al., 2021).

Until now, SGM has been applied to Euclidean data, i.e. data with flat geometry. However, in a large number of scientific domains, the underlying assumption is that the distributions of interest are supported on a Riemannian manifold. These include, amongst others, protein modeling (Boomsma et al., 2008; Hamelryck et al., 2006; Mardia et al., 2008; Shapovalov and Dunbrack Jr, 2011; Mardia et al., 2007), cell development (Klimovskaia et al., 2020), image recognition (Lui, 2012), geological sciences (Karpatne et al., 2018; Peel et al., 2001), graph-structured and hierarchical data (Roy et al., 2007; Steyvers and Tenenbaum, 2005), robotics (Feiten et al., 2013; Senanayake and Ramos, 2018) and high-energy physics (Brehmer and Cranmer, 2020). The choice of a Riemannian metric is associated with a description of the interactions between the points of the dataset and therefore can be seen as a geometric prior.

In this paper we introduce *Riemannian Score-based Generative Models* (RSGM), an extension of SGMs to compact Riemannian manifolds. Contrary to classical SGMs which rely on forward and time-reversed diffusion processes defined on an Euclidean space, we incorporate the geometry of the data in our algorithm by defining our diffusion processes directly on the Riemannian manifold. However, switching from the classical Euclidean setting to the Riemannian is non-trivial. First, one must be able to define a noising process on the manifold that converges to an easy-to-sample reference distribution. In the setting of compact Riemannian manifolds, a natural choice is given by the Brownian motion. Indeed, due to the compactness, this diffusion is geometrically ergodic and targets the uniform distribution on the manifold (He, 2013) from which one can either sample exactly or approximately with high accuracy. Second, we must identify the corresponding time-reversal process. We show here that, as in the Euclidean case, this process is also a diffusion whose infinitesimal generator is given by the generator of the forward process with an extra term corresponding to the scores of the marginal distributions of the Brownian diffusion initialized at the data distribution. Third, while score matching ideas (Hyvärinen and Dayan, 2005; Vincent, 2011) can be easily used to estimate the score in the Euclidean case when the forward dynamics is given by a Ornstein–Ulhenbeck or a Brownian motion, adapting these ideas to the Riemmanian framework is complicated by the fact that the heat kernel, i.e. the transition kernel of the Brownian motion, is typically only available as an infinite sum through the Sturm-Liouville decomposition. Similarly, diffusions on manifold cannot be sampled exactly. Hence, we use geodesic random walks which converge to the diffusion of interest in the limit of small stepsizes (Jørgensen, 1975).

We further consider the following extensions of RSGMs. By using tools from neural ODEs on manifolds (Mathieu and Nickel, 2020; Falorsi and Forré, 2020; Lou et al., 2020), we show how we can compute the likelihood of our model, generalizing the approach proposed in the Euclidean case in (Song et al., 2021; Durkan and Song, 2021; Huang et al., 2021). Finally, RGSMs like standard SGMs are computationally expensive at generation time as they require to run a discretized diffusion over many time steps. For speeding up generation, it has been proposed in the Euclidean setting to solve instead a Schrödinger Bridge (SB) problem (De Bortoli et al., 2021; Chen et al., 2021a), i.e. a dynamical version of an entropy-regularized Optimal Transport (OT) problem between the data and the easy-to-sample reference distribution. In particular, we generalize the Diffusion Schrödinger Bridge (DSB) algorithm introduced in (De Bortoli et al., 2021) to solve the SB problem on compact Riemmanian manifolds.

We validate our methodology by modelling a number of natural disaster occurrence datasets collected by Mathieu and Nickel (2020). We compare to three previous baselines, a mixture of Kent distributions Peel et al. (2001), Riemannian Continuous Normalising Flows Mathieu and Nickel (2020), and Moser Flows Rozen et al. (2021). We also compare to using a standard standard Euclidean SGM by projecting the manifold onto Euclidean space and performing the flow there (e.g. projecting the sphere via the stereographic projection onto the 2D plane). We find in all cases that RSGMs outperform all baselines.

The rest of the paper is organized as follows. We introduce the notation needed in the rest of the paper in Section 2. We recall the basics of standard Euclidean SGMs in Section 3. In Section 4, we present RGSMs, our extension of SGMs to compact Riemannian manifolds. We discuss related works in Section 5 and assess the efficiency of our method in Section 6. Finally we summarize our contributions in Section 7.

2 Notation

We consider a compact connected Riemannian manifold $(\mathcal{M}, \langle \cdot, \cdot \rangle_{\mathcal{M}})$. We denote by $\mathcal{X}(\mathcal{M})$ the set of vector fields on $\mathcal{M}$ and $\mathcal{X}^2(\mathcal{M})$ the section $\Gamma(\mathcal{M}, \sqcup_{x \in \mathcal{M}} \mathcal{L}(\mathrm{T}_x\mathcal{M}))$, where $\mathcal{L}(\mathrm{T}_x\mathcal{M})$ is the space of linear mappings on $\mathrm{T}_x\mathcal{M}$. Let $(\mathbf{M}_t)_{t\in[0,T]}$ be a real-valued process and $(\mathbf{X}_t)_{t\in[0,T]}$ be a $\mathcal{M}$-valued process with distribution $\mathbb{P} \in \mathcal{P}(\mathrm{C}([0,T], \mathcal{M}))$. $(\mathbf{M}_t)_{t\in[0,T]}$ is a $\mathbb{P}$-martingale if $(\mathbf{M}_t)_{t\in[0,T]}$ is a martingale w.r.t the filtration $(\mathcal{F}_t)_{t\in[0,T]}$ where for any $t \in [0,T]$, $\mathcal{F}_t = \sigma(\{\mathbf{X}_s : s \in [0,t]\})$. In addition, for any $\mathbb{P} \in \mathcal{P}(\mathrm{C}([0,T], \mathcal{M}))$, we define $R(\mathbb{P})$ such that for any $\mathsf{A} \in \mathcal{B}(\mathrm{C}([0,T], \mathsf{X}))$ we have $R(\mathbb{P})(\mathsf{A}) = \mathbb{P}(R(\mathsf{A}))$, where $R(\mathsf{A}) = \{t \mapsto \omega_{T-t} : \omega \in \mathsf{A}\}$. In other words, $R(\mathbb{P})$ is the path measure associated with the reverse process $\mathbb{P}$. When there is no ambiguity, we use the same notation for distributions and their densities.

Let $T > 0$ or $T = +\infty$, $b : [0,T] \to \mathcal{X}(\mathcal{M})$, $\Sigma : [0,T] \to \mathcal{X}^2(\mathcal{M})$ such that for any $t \in [0,T]$ and $x \in \mathcal{M}$, $\Sigma(t,x)$ is symmetric, non-negative and denote $\sigma(t,x) = \Sigma^{1/2}(t,x)$. Let $(\mathbf{X}_t)_{t\in[0,T]}$ a continuous process with distribution $\mathbb{P} \in \mathcal{P}(\mathrm{C}([0,T], \mathcal{M}))$ such that for any $f \in \mathrm{C}^2(\mathcal{M})$ we have that $(\mathbf{M}_t^{\mathbf{X},f})_{t\in[0,T]}$ is a $\mathbb{P}$-martingale where for any $t \in [0,T]$

$$\mathbf{M}_t^{\mathbf{X},f} = f(\mathbf{X}_t) - \int_0^t \{\langle b(s, \mathbf{X}_s), \nabla f(\mathbf{X}_s)\rangle_{\mathcal{M}} + (1/2)\langle \Sigma(\mathbf{X}_s), \nabla^2 f(\mathbf{X}_s)\rangle_{\mathcal{M}}\}\mathrm{d}s.$$

Then, we say that $(\mathbf{X}_t)_{t\in[0,T]}$ is *associated with* the SDE $\mathrm{d}\mathbf{X}_t = b(t, \mathbf{X}_t)\mathrm{d}t + \sigma(t, \mathbf{X}_t)\mathrm{d}\mathbf{B}_t^{\mathcal{M}}$ with infinitesimal generator $\mathcal{A} : [0,T] \times \mathrm{C}^2(\mathcal{M}) \to \mathrm{C}(\mathcal{M})$ given for any $t \in [0,T]$ by $\mathcal{A}_t(f) = \langle b, \nabla f\rangle_{\mathcal{M}} + (1/2)\langle \Sigma, \nabla^2 f\rangle_{\mathcal{M}}$. Note that if $\Sigma = \mathrm{Id}$ then $\langle \Sigma, \nabla^2 f\rangle_{\mathcal{M}} = \Delta f$, where Δ is the Laplace-Beltrami operator.

3 Euclidean Score-based Generative Modeling

We recall here briefly the key concepts behind SGMs on the Euclidean space $\mathbb{R}^d$ for some $d \in \mathbb{N}$. We refer to Song et al. (2021); Song and Ermon (2019); De Bortoli et al. (2021) for a more detailed introduction to SGMs. In what follows, let p_0 denote the data distribution. We have practically only access to an empirical approximation of this distribution given by the available data.

We consider a forward *noising* process $(\mathbf{X}_t)_{t\geq 0}$ defined by the following Stochastic Differential Equation (SDE)

$$\mathrm{d}\mathbf{X}_t = -\mathbf{X}_t \mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t, \quad \mathbf{X}_0 \sim p_0 \tag{1}$$

where $(\mathbf{B}_t)_{t\geq 0}$ is a d-dimensional Brownian motion. As a result $(\mathbf{X}_t)_{t\geq 0}$ is an Ornstein–Ulhenbeck process targeting a multivariate standard Gaussian distribution. Let $T \geq 0$, under mild conditions on the data distribution p_0, the time-reversed process $(\hat{\mathbf{X}}_t)_{t\geq 0} = (\mathbf{X}_{T-t})_{t\in[0,T]}$ also satisfies an SDE (Cattiaux et al., 2021; Haussmann and Pardoux, 1986) given by

$$\mathrm{d}\hat{\mathbf{X}}_t = \{\hat{\mathbf{X}}_t + 2\nabla \log p_{T-t}(\hat{\mathbf{X}}_t)\}\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t, \quad \hat{\mathbf{X}}_0 \sim p_T \tag{2}$$

where p_t denotes the density of $\mathbf{X}_t$. By construction, the law of $\hat{\mathbf{X}}_{T-t}$ is equal to the law of $\mathbf{X}_t$ for $t \in [0, T]$ and in particular $\hat{\mathbf{X}}_T \sim p_0$. Hence, if one could sample from $(\hat{\mathbf{X}}_t)_{t\in[0,T]}$ then its final distribution would be the target data distribution p_0.

Unfortunately there are three sources of intractability that prevents us from sampling the process $(\hat{\mathbf{X}}_t)_{t\in[0,T]}$.

Problem 1: Its initial distribution is given by p_T which is intractable.

Solution: The Ornstein–Ulhenbeck process (1) converges exponentially fast towards a standard multivariate Gaussian so one can approximate p_T by this Gaussian for T large enough.

Problem 2: The scores are intractable so the dynamics (2) cannot be implemented.

Solution: To approximate the scores, we exploit the following identity

$$\nabla \log p_t(x) = \int_{\mathbb{R}^d} \nabla \log p_{t|0}(x|x_0) p_{0|t}(x_0|x)\mathrm{d}x_0,$$

where $p_{t|0}(x'|x)$ is the transition density of the Ornstein–Ulhenbeck process which is available in closed-form. It follows directly that $\nabla \log p_t$ is the minimizer of the loss function $\ell_t(s) = \mathbb{E}[\|s(\mathbf{X}_t) - \nabla \log p_{t|0}(\mathbf{X}_t|\mathbf{X}_0)\|^2]$ over function s where the expectation is over the joint distribution of $\mathbf{X}_0, \mathbf{X}_t$. This result can be exploited as follows. We consider a neural network approximation $\boldsymbol{s}_\theta : [0, T] \times \mathbb{R}^d \to \mathbb{R}^d$ which we train by minimizing the loss function $\ell(\theta) = \int_0^T \lambda_t \ell_t(\boldsymbol{s}_\theta(t, \cdot))\mathrm{d}t$ for some weighting function $\lambda_t > 0$.

Problem 3: The loss function $\ell(\theta)$ and the SDE approximating (2) by replacing the scores $(\nabla \log p_t)_{t\in[0,T]}$ by $(s_\theta(t, \cdot))_{t\in[0,T]}$ and p_T by the standard multivariate normal cannot not be simulated exactly on a computer.

Solution: For a discretization step γ such that $T = \gamma N$ for integer N, the loss function is approximated by $\sum_{n=0}^N \lambda_{n\gamma} \ell_{n\gamma}(\boldsymbol{s}_\theta(n\gamma, \cdot))$ and we perform a Euler–Maruyama discretization of the resulting SDEe; i.e. we define $(Y_n)_{n\in\{0,\ldots,N\}}$ such that for $Z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$

$$Y_{n+1} = Y_n + \gamma\{Y_n + 2\boldsymbol{s}_\theta(T - n\gamma, Y_n)\} + \sqrt{2}Z_{n+1}, \quad Y_0 \sim \mathcal{N}(0, I_d).$$

We have presented the basics of SGM but we highlight that many recent works improve on these models; (see e.g. Song et al., 2021; Song and Ermon, 2020; Song et al., 2020; Jolicoeur-Martineau et al., 2021b;a; Dhariwal and Nichol, 2021). In particular, it is worth noting that choosing an adaptive stepsize $(\gamma_n)_{n\in\mathbb{N}}$ (Bao et al., 2022; Watson et al., 2021) drastically improve the synthesis results as well as using a predictor-corrector scheme (Song et al., 2021) instead of a simple Euler–Maruyama discretization. Finally, we note that there exist other approaches to introduce SGMs using variational and maximum likelihood techniques (Ho et al., 2020; Huang et al., 2021; Durkan and Song, 2021).

4 Riemannian Score-based Generative Modeling

Similarly to the Euclidean setting, three ingredients are required to extend SGM to compact Riemannian manifolds: i) a forward *noising* process on the Riemannian manifold which converges to an easy-to-sample reference distribution, ii) a time-reversal formula on Riemannian manifolds which defines a backward generative process, iii) a method to efficiently approximate the drift of the time-reversal process. We address all these problems and introduce RGSM. The key differences between SGMs and RSGMs are summarised in Table 1.

Ingredient \ Space	Euclidean	Compact manifold
Forward process	Ornstein–Ulhenbeck	Brownian motion
Easy-to-sample distribution	Gaussian	Uniform
Time reversal	Cattiaux et al. (2021, Theorem 4.9)	Theorem 5
Sampling of the forward process	Direct	Geodesic Random Walk
Sampling of the backward process	Euler–Maruyama	Geodesic Random Walk

Table 1: Differences between SGM on Euclidean spaces and RSGM on compact Riemannian manifolds.

4.1 Brownian motion on compact Riemannian manifolds

Brownian motion and uniform distribution First, we define a forward noising process on $\mathcal{M}$ targeting an easy-to-sample reference distribution. In Euclidean spaces, the reference distribution is a standard normal in the compact manifold setting the uniform distribution p_{ref} is the natural choice. For most manifolds of interest, one can either sample exactly from p_{ref} or obtain approximate samples with high accuracy. For the forward noising dynamics, the Ornstein–Ulhenbeck process (1) used in Euclidean scenarios is now replaced by the Brownian motion defined on $\mathcal{M}$ as it converges exponentially fast to p_{ref}—see Proposition 2 below. We refer to Appendix S2.5 for a general introduction to Brownian motions on manifolds. This Brownian motion is defined as follows.

Definition 1 (Brownian motion)**.** *$(\mathbf{B}_t^{\mathcal{M}})_{t\geq 0}$ is a Brownian motion on $\mathcal{M}$ if $(\mathbf{B}_t^{\mathcal{M}})_{t\geq 0}$ is associated with the SDE with infinitesimal generator $\mathcal{A}(f) = \Delta f$, see Section 2.*

We refer to Appendix S2.5 or Hsu (2002, Chapter 1, Chapter 3) for the definition of a $\mathcal{M}$-valued semimartingale and the Laplace-Beltrami operator. By Hsu (2002, Proposition 3.2.1), we have that for any initial condition $\mathbf{B}_0^{\mathcal{M}}$ with distribution $p_0 \in \mathcal{P}(\mathcal{M})$, there exists $(\mathbf{B}_t^{\mathcal{M}})_{t\geq 0}$. The convergence rates are obtained w.r.t. the total variation distance between the uniform distribution and the semi-group $(\mathrm{P}_t)_{t\geq 0}$[1]

Proposition 2 (Convergence of Brownian motion (Urakawa, 2006, Proposition 2.6))**.** *For any $t > 0$, P_t admits a density p_t w.r.t p_{ref} and $p_{\mathrm{ref}}\mathrm{P}_t = p_{\mathrm{ref}}$, i.e. p_{ref} is an invariant measure for $(\mathrm{P}_t)_{t\geq 0}$. In addition, if there exists $C, \alpha \geq 0$ such that for any $t \in (0, 1]$, $p_t(x, x) \leq Ct^{-\alpha/2}$ then for any $p_0 \in \mathcal{P}(\mathcal{M})$ and for any $t \geq 1/2$ we have*

$$\|p_0\mathrm{P}_t - p_{\mathrm{ref}}\|_{\mathrm{TV}} \leq C^{1/2}\mathrm{e}^{\lambda_1/2}\mathrm{e}^{-\lambda_1 t},$$

where λ_1 is the first non-negative eigenvalue of $-\Delta_{\mathcal{M}}$ in $\mathrm{L}^2(p_{\mathrm{ref}})$.

The diagonal upper bound on the heat kernel is satisfied for many manifolds including the d-dimensional torus and sphere (see Saloff-Coste, 1994, Section 3). Hence, Proposition 2 ensures that under mild conditions the Brownian motion converges exponentially fast towards the uniform distribution on the compact Riemmanian manifold $\mathcal{M}$. Therefore, in the context of SGM, the Brownian motion on $\mathcal{M}$ is the counterpart to the Ornstein–Ulhenbeck process and the uniform distribution is the counterpart to the Gaussian one.

We note that in previous works on SGMs, the Brownian motion has also been used as a forward noising process (Song et al., 2021; Song and Ermon, 2019). However, in these cases, the Brownian motion is not geometrically ergodic and does not admit any invariant distribution contrary to our setting. Two issues remain to be solved. First, we need to be able to sample this forward noising process $(\mathbf{B}_t^{\mathcal{M}})_{t\geq 0}$. Second we need to obtain tractable approximations of the heat kernel, i.e. the transition kernel of this process, in order to define efficient score approximation schemes.

Sampling from diffusions In Euclidean spaces, sampling an Ornstein–Ulhenbeck process is straightforward whereas obtaining samples from a Brownian motion on a manifold is non-trivial in general. First, if $\mathcal{M}$ is isometrically embedded into $\mathbb{R}^p$ (with $p \geq d$)—i.e. $\mathcal{M} \subset \mathbb{R}^p$—then we have that $(\mathbf{B}_t^{\mathcal{M}})_{t\geq 0}$ (seen as a process on the ambient space $\mathbb{R}^p$) satisfies the following SDE

$$\mathrm{d}\mathbf{B}_t^{\mathcal{M}} = \textstyle\sum_{i=1}^p P_i(\mathbf{B}_t^{\mathcal{M}}) \circ \mathrm{d}\mathbf{B}_t^i,$$

[1]We define $(\mathrm{P}_{t,s})_{t,s\geq 0, t\geq s}$ the semi-group such that for any $f, g \in \mathrm{C}(\mathcal{M})$ and $t, s \geq 0$ with $t \geq s$ we have $\mathbb{E}[f(\mathbf{B}_t^{\mathcal{M}})g(\mathbf{B}_s^{\mathcal{M}})] = \mathbb{E}[\int_{\mathcal{M}} f(y)\mathrm{P}_{t|s}(\mathbf{B}_s^{\mathcal{M}}, \mathrm{d}y)g(\mathbf{B}_s^{\mathcal{M}})]$. For the rest of this paper we denote $(\mathrm{P}_t)_{t\geq 0} = (\mathrm{P}_{t|0})_{t\geq 0}$.

Algorithm 1 Geodesic Random Walk (GRW)

Require: T, K, X_0, b, σ, P
1: $\gamma = T/K$ ▷ Step-size
2: **for** $k \in \{0, \dots, K-1\}$ **do**
3: $\quad \bar{Z}_{k+1} \sim \mathcal{N}(0, I_p)$ ▷ Standard Gaussian in ambient space $\mathbb{R}^p$
4: $\quad Z_{k+1} = P(X_k)\bar{Z}_{k+1}$ ▷ Projection in the tangent space $\mathrm{T}_x\mathcal{M}$
5: $\quad V_{k+1} = \gamma b(k\gamma, X_k) + \sqrt{\gamma}\sigma(k\gamma, X_k)Z_{k+1}$ ▷ Euler-Maruyama step on tangent space
6: $\quad X_{k+1} = \exp_{X_k}(V_{k+1})$ ▷ Geodesic projection onto $\mathcal{M}$
7: **end for**
8: **return** $\{X_k\}_{k=0}^{K-1}$

where $\circ$ denotes the Stratanovitch integral [2], $(\{\mathbf{B}_t^i\}_{i=1}^p)_{t\geq 0}$ is a p-dimensional Brownian motion and for any $i \in \{1, \dots, p\}$ we have $P_i(x) = P(x)e_i$ for any $x \in \mathcal{M}$, where $\{e_i\}_{i=1}^p$ is the canonical basis of $\mathbb{R}^p$ and $P(x): \mathbb{R}^p \to \mathrm{T}_x\mathcal{M}$ is the orthogonal projection operator, see Appendix S2.1. However, this approach is *extrinsic* and requires the knowledge of the projection operator. Here we consider an *intrisic* approach based on Geodesic Random Walks (GRWs), see Jørgensen (1975) for a review of their properties.

GRWs are not restricted to approximating the Brownian motion on $\mathcal{M}$ but in fact can approximate *any* well-behaved diffusion on $\mathcal{M}$. This property will be useful when sampling the backward diffusion process. Hence, we introduce GRWs in a general framework and we are going to consider a discrete-time process $(X_n^\gamma)_{n\in\mathbb{N}}$ which approximates $(\mathbf{X}_t)_{t\geq 0}$ is associated with

$$\mathrm{d}\mathbf{X}_t = b(\mathbf{X}_t)\mathrm{d}t + \sigma(\mathbf{X}_t)\mathrm{d}\mathbf{B}_t^{\mathcal{M}}. \tag{3}$$

Let $\{\nu_x\}_{x\in\mathcal{M}}$ such that for any $x \in \mathcal{M}$, $\nu_x \in \mathcal{P}(\mathrm{T}_x\mathcal{M})$. Assume that for any $x \in \mathcal{M}$, $\int_{\mathcal{M}} \|v\|^2 \mathrm{d}\nu_x(v) < +\infty$. In addition assume that there exists $b \in \mathcal{X}(\mathcal{M})$ and $\Sigma \in \mathcal{X}^2(\mathcal{M})$, such that for any $x \in \mathcal{M}$, $\int_{\mathcal{M}} v\mathrm{d}\nu_x(v) = b(x)$ and $\int_{\mathcal{M}}(v - b(x)) \otimes (v - b(x))\mathrm{d}\nu_x(v) = \Sigma(x) = \sigma(x)\sigma(x)^\top$.

Definition 3 (Geodesic Random Walk)**.** *Let X_0 be a $\mathcal{M}$-valued random variable. For any $\gamma > 0$, we define $(\mathbf{X}_t^\gamma)_{t\geq 0}$ such that $\mathbf{X}_0^\gamma = X_0$ and for any $n \in \mathbb{N}$ and $t \in [0, \gamma]$, $\mathbf{X}_{n\gamma+t} = \exp_{\mathbf{X}_{n\gamma}}(t\gamma\{\mu_n + (1/\sqrt{\gamma})(V_n - \mu_n)\})$[3], where $(V_n)_{n\in\mathbb{N}}$ is a sequence of random variables in such that for any $n \in \mathbb{N}$, V_n has distribution $\nu_{\mathbf{X}_{n\gamma}}$ conditionally to $\mathbf{X}_{n\gamma}$. We say that $(X_n^\gamma)_{n\in\mathbb{N}} = (\mathbf{X}_{n\gamma})_{n\in\mathbb{N}} \in \mathcal{M}$ is a Geodesic Random Walk (GRW).*

Note that for any $n \in \mathbb{N}$ and $\gamma > 0$, $X_n^\gamma \in \mathcal{M}$. For any $\gamma > 0$, we denote by $(\mathrm{Q}_n^\gamma)_{n\in\mathbb{N}}$ the sequence of Markov kernels such that for any $n \in \mathbb{N}$, $x \in \mathcal{M}$ and $\mathsf{A} \in \mathcal{B}(\mathcal{M})$ we have that $\delta_x\mathrm{Q}_n^\gamma(\mathsf{A}) = \mathbb{P}(X_n^\gamma \in \mathsf{A})$, with $X_0^\gamma = x$. GRWs are appealing because, under mild conditions, when the stepsize $\gamma \to 0$, they converge towards $(\mathbf{X}_t)_{t\geq 0}$ solution of (3) in the following sense:

Theorem 4 (Convergence of geodesic random walk (Jørgensen, 1975, Theorem 2.1))**.** *Under the conditions of Theorem S14, for any $t \geq 0$, $f \in \mathrm{C}(\mathcal{M})$ we have that $\lim_{\gamma\to 0} \|\mathrm{Q}_\gamma^{\lceil t/\gamma\rceil}[f] - \mathrm{P}_t[f]\|_\infty = 0$, where $(\mathrm{P}_t)_{t\geq 0}$ is the semi-group associated with the infinitesimal generator $\mathcal{A}: \mathrm{C}^\infty(\mathcal{M}) \to \mathrm{C}^\infty(\mathcal{M})$ given for any $f \in \mathrm{C}^\infty(\mathcal{M})$ by $\mathcal{A}(f) = \langle b, \nabla f\rangle_{\mathcal{M}} + (1/2)\langle\Sigma, \nabla^2 f\rangle_{\mathcal{M}}$.*

In particular if $b = 0$ and $\sigma = \mathrm{Id}$, then the random walk converges towards a Brownian motion on $\mathcal{M}$ in the sense of the convergence of semi-groups. One advantage of GRW is that they allow to samples from arbitrary diffusions under mild assumptions. This property will be key to sample from the backward process. Theorem 4 can be extended to approximate time inhomogeneous diffusions. We leave the proof of this extension for future work. In Algorithm 1, we remind how to approximately sample from a diffusion $(\mathbf{X}_t)_{t\in[0,T]}$ using GRWs, where $(\mathbf{X}_t)_{t\in[0,T]}$ associated with the family of infinitesimal generator $(\mathcal{A}_t)_{t\in[0,T]}$ given for any $t \in [0, T]$ and $f \in \mathrm{C}^2(\mathcal{M})$ by $\mathcal{A}_t(f) = \langle b_t, \nabla f\rangle + \langle\Sigma_t, \nabla^2 f\rangle$, where $b: [0, T] \to \mathcal{X}(\mathcal{M})$, $\Sigma_t = \sigma_t\sigma_t^\top$ with $\sigma_t: [0, T] \to \mathcal{X}^2(\mathcal{M})$. For simplicity, in Algorithm 1, we assume that $\mathcal{M}$ is embedded in $\mathbb{R}^p$ and use the projection to define the noise on the tangent space (such an embedding always exists using the Nash embedding theorem, see Gunther (1991) for example). In a more general setting, we compute the noise on the tangent space using local coordinates.

[2] Manifold valued processes are usually defined using the Stratanovitch integral because it satisfies the chain rule of differential calculus. For more details we refer to Appendix S2.2.

[3] where $\exp_x: \mathrm{T}_x\mathcal{M} \to \mathcal{M}$ is the exponential mapping on the manifold, see Lee (2013, Chapter 20) for details.

Heat kernel on compact Riemannian manifolds The semi-group of the Brownian motion $(\mathrm{P}_t)_{t\geq 0}$ (called the heat kernel) admits a density w.r.t. p_{ref}, such that for any $f \in \mathrm{C}(\mathcal{M})$, $x_0 \in \mathcal{M}$ and $t > 0$ we have

$$\delta_{x_0}\mathrm{P}[f] = \int_{\mathcal{M}} f(x_t) p_{t|0}(x_t|x_0) \mathrm{d}p_{\mathrm{ref}}(x_t).$$

In addition, this transition density is positive and $(t, x, y) \mapsto p_{t|0}(y|x) \in \mathrm{C}^\infty((0, +\infty) \times \mathcal{M} \times \mathcal{M})$ and satisfies the heat equation $\partial_t p_{t|0} = \Delta p_{t|0}$. However, contrary to the Gaussian transition density of the Ornstein–Ulhenbeck process, it is typically only available as an infinite series. In order to circumvent this issue we consider two techniques: i) a truncation approach, ii) a Taylor expansion around $t = 0$, *i.e.* a Varadhan asymptotics. First, we recall that in the case of compact manifolds we have that for any $t > 0$ and $x, y \in \mathcal{M}$

$$p_{t|0}(x, y) = \sum_{j\in\mathbb{N}} \mathrm{e}^{-\lambda_j t} \phi_j(x)\phi_j(y), \tag{4}$$

where the convergence occurs in $\mathrm{L}^2(p_{\mathrm{ref}} \otimes p_{\mathrm{ref}})$, $(\lambda_j)_{j\in\mathbb{N}}$ and $(\phi_j)_{j\in\mathbb{N}}$ are the eigenvalues, respectively the eigenvectors, of $-\Delta_{\mathcal{M}}$ in $\mathrm{L}^2(p_{\mathrm{ref}})$ (see Saloff-Coste, 1994, Section 2). When the eigenvalues and eigenvectors are known, we approximate the logarithmic gradient of $p_{t|0}$ by truncating the sum in Equation (4) with $J \in \mathbb{N}$ terms to obtain for any $t > 0$ and $x, y \in \mathcal{M}$

$$\nabla_x \log p_{t|0}(x, y) \approx S_{J,t}(x, y) = \sum_{j=0}^{J} \mathrm{e}^{-\lambda_j t} \nabla\phi_j(x)\phi_j(y) / \sum_{j=0}^{J} \mathrm{e}^{-\lambda_j t}\phi_j(x)\phi_j(y).$$

Note that for any $t \geq 0$, $x, y \in \mathcal{M}$, $S_{J,t}(x, y) \in \mathrm{T}_x\mathcal{M}$. Under regularity conditions on $\mathcal{M}$ it can be shown that for any $x, y \in \mathcal{M}$ and $t \geq 0$, $\lim_{J\to+\infty} S_{J,t} = \nabla_x \log p_{t|0}(x, y)$ (see Jones et al., 2008, Lemma 1). In the case of the d-dimensional torus or sphere the eigenvalues and eigenvectors are known, (see Saloff-Coste, 1994, Section 2) and we can apply this method to approximate $p_{t|0}$ for any $t > 0$. We refer to Appendix S4 for more details about eigenvalues and eigenfunctions of the Laplace-Beltrami operator in the special case of the d-dimensional torus and sphere.

When the eigenvalues and eigenvectors are not tractable, we can still derive an approximation of the heat kernel for small times t. Using Varadhan's asymptotics—see Bismut (1984, Theorem 3.8) or Chen et al. (2021b, Theorem 2.1)—for any $x, y \in \mathcal{M}$ with $y \notin \mathrm{Cut}(x)$ (where $\mathrm{Cut}(x)$ is the cut-locus of x in $\mathcal{M}$) we have that (see Lee, 2018, Chapter 10)

$$\lim_{t\to 0} t\nabla_y \log p_{t|0}(x, y) = -\exp_x^{-1}(y). \tag{5}$$

VDB: experiment: quality o the approximation as a function of J and t. On the same graph put the Varadhan approx

Note that since the cut-locus has null measure under the uniform distribution p_{ref} (Lee, 2006, Theorem 10.34), the previous relation is valid almost everywhere. We will see in Section 4.3 that an approximation for any $x, y \in \mathcal{M}$ of $t\nabla_y \log p_t(x, y)$ for small values of $t \geq 0$ is enough to define a score approximation.

4.2 A manifold time-reversal formula

After having defined the forward noising process targeting a reference distribution, a second key ingredient of SGMs is to derive a time-reversal formula. Namely, if $(\mathbf{X}_t)_{t\in[0,T]}$ is a diffusion process then $(\mathbf{X}_{T-t})_{t\in[0,T]}$ is also a diffusion process w.r.t. the backward filtration whose coefficients can be computed, see Appendix S5. Our next result is the Riemannian counterpart to the Euclidean time-reversal formula, see Cattiaux et al. (2021, Theorem 4.9) and Haussmann and Pardoux (1986) for instance, which states under mild regularity and integrability conditions if the $\mathbb{R}^d$-valued process $(\mathbf{X}_t)_{t\in[0,T]}$ is a (weak) solution to the SDE

$$\mathrm{d}\mathbf{X}_t = b(\mathbf{X}_t)\mathrm{d}t + \mathrm{d}\mathbf{B}_t,$$

then $(\mathbf{Y}_t)_{t\in[0,T]} = (\mathbf{X}_{T-t})_{t\in[0,T]}$ is a (weak) solution to the SDE

$$\mathrm{d}\mathbf{Y}_t = \{-b(\mathbf{Y}_t) + \nabla \log p_{T-t}(\mathbf{Y}_t)\}\mathrm{d}t + \mathrm{d}\mathbf{B}_t,$$

In the case where $(\mathbf{X}_t)_{t\in[0,T]}$ is an Ornstein-Ulhenbeck process, then we recover Equation (2).

Theorem 5 (Reverse diffusion). *Let $T \geq 0$ and $(\mathbf{B}_t^{\mathcal{M}})_{t\geq 0}$ be a Brownian motion on $\mathcal{M}$ such that $\mathbf{B}_0^{\mathcal{M}}$ has distribution p_{ref}. Let $(\mathbf{X}_t)_{t\in[0,T]}$ associated with the SDE $\mathrm{d}\mathbf{X}_t = b(\mathbf{X}_t)\mathrm{d}t + \mathrm{d}\mathbf{B}_t^{\mathcal{M}}$. Let $(\mathbf{Y}_t)_{t\in[0,T]} = (\mathbf{X}_{T-t})_{t\in[0,T]}$ and assume that $\mathrm{KL}(\mathbb{P}|\mathbb{Q}) < +\infty$, where $\mathbb{Q} \in \mathcal{P}(\mathrm{C}([0,T], \mathcal{M}))$ is the distribution of $(\mathbf{B}_t^{\mathcal{M}})_{t\in[0,T]}$. In addition, assume that for any $t \in [0,T]$, $\mathbb{P}_t$ admits a smooth positive density p_t w.r.t. p_{ref}. Then, we have that $(\mathbf{Y}_t)_{t\in[0,T]}$ is associated with the SDE*

$$\mathrm{d}\mathbf{Y}_t = \{-b(\mathbf{Y}_t) + \nabla \log p_{T-t}(\mathbf{Y}_t)\}\mathrm{d}t + \mathrm{d}\mathbf{B}_t^{\mathcal{M}}. \tag{6}$$

Proof. The proof is a smooth extension of Cattiaux et al. (2021, Theorem 4.9) to the Riemannian manifold case. We postpone the detailed proof to Appendix S5. □

Note the formula obtained for the drift of the time reversed process are the same in the Euclidean and the Riemannian settings upon replacing the Euclidean gradient operator, Laplacian and scalar product by the Riemannian ones. As a corollary of Theorem 5, we get the following result.

Corollary 6. *Under the conditions of Theorem 5, denote* $(\mathrm{P}_t)_{t\in[0,T]}$ *and* $(\mathrm{Q}_{s,t})_{s,t\in[0,T],t\geq s}$ *the semi-group associated with* $\mathbb{P}$*, respectively* $R(\mathbb{P})$*. Then:*

(a) $(\mathrm{P}_t)_{t\in[0,T]}$ *is associated with the generator* $\mathcal{A} : \mathrm{C}^\infty(\mathcal{M}) \to \mathrm{C}^\infty(\mathcal{M})$ *given for any* $f \in \mathrm{C}^\infty(\mathcal{M})$ *by* $\mathcal{A}(f) = \langle b, \nabla f\rangle_{\mathcal{M}} + (1/2)\Delta_{\mathcal{M}} f$*.*

(b) $(\mathrm{Q}_{t|s})_{s,t\in[0,T],t\geq s}$ *is associated with the family of generators* $(\mathcal{A}_u)_{u\in[0,T]}$ *such that for any* $u \in [0,T]$*,* $\mathcal{A}_u : \mathrm{C}^\infty(\mathcal{M}) \to \mathrm{C}^\infty(\mathcal{M})$ *is given for any* $f \in \mathrm{C}^\infty(\mathcal{M})$ *by* $\mathcal{A}_u(f) = \langle -b + \nabla \log p_{T-u}, \nabla f\rangle_{\mathcal{M}} + (1/2)\Delta_{\mathcal{M}} f$*.*

In particular, we can approximate the time-reversed process using a GRW using Theorem 4.

4.3 Score approximation on Riemannian manifolds

The last ingredient in order to define the (compact) Riemannian manifold extension of SGM is an approximation of the logarithmic gradient appearing in Equation (6).

Score-matching and loss functions We aim to approximate $\nabla \log p_t(x)$ for every $t \in (0,T]$ and $x \in \mathcal{M}$. To do so, we first remark that for any $s,t \in (0,T]$ with $t > s$ and $x_t \in \mathcal{M}$, $p_t(x_t) = \int_{\mathcal{M}} p_{t|s}(x_t|x_s)\mathrm{d}\mathbb{P}_s(x_s)$. Therefore, we obtain that for any $s,t \in [0,T]$ with $t > s$ and $x_t \in \mathcal{M}$

$$\nabla \log p_t(x_t) = \int_{\mathcal{M}} \nabla_x \log p_{t|s}(x_t|x_s)\mathrm{P}_{s|t}(x_t, \mathrm{d}x_s).$$

Hence, for any $s,t \in [0,T]$ with $t > s$ we have that

$$\nabla \log p_t = \arg\min\{\ell_{t|s}(s_t) \,:\, s_t \in \mathrm{L}^2(\mathbb{P}_t)\}, \quad \ell_{t|s}(s_t) = \int_{\mathcal{M}\times\mathcal{M}} \|\nabla_x \log p_{t|s}(x_t|x_s) - s_t(x_t)\|^2 \mathrm{d}\mathbb{P}_{s,t}(x_s,x_t).$$

The loss function $\ell_{t|s}$ is called the Denoising Score Matching (DSM) loss. It can also be written in an *implicit* fashion.

Proposition 7. *Let* $t \in (0,T]$*. If* $s_t \in \mathrm{C}^\infty(\mathcal{M})$ *then we have that* $\ell_{t|s}(s_t) = 2\ell_t^{\mathrm{im}}(s_t) + \int_{\mathcal{M}\times\mathcal{M}} \|\nabla \log p_{t|s}(x_t|x_s)\|^2 \mathrm{d}\mathbb{P}_{s,t}(x_s,x_t)$*, where*

$$\ell_t^{\mathrm{im}}(s_t) = \int_{\mathcal{M}} \{\tfrac{1}{2}\|s_t(x_t)\|^2 + \mathrm{div}(s_t)(x_t)\}\mathrm{d}\mathbb{P}_t(x_t).$$

Proof. The proof is postponed to Appendix S7. □

For any $t \in (0,T]$ the minimizers of the loss ℓ_t^{im} on $\mathcal{X}(\mathcal{M})$ are the same as the ones for $\ell_{t|s}$. The loss ℓ_t^{im} is called the *implicit* score matching (ISM) loss (or sliced score matching (SSM) loss if the divergence is approximated using the Hutchinson's trace estimator Hutchinson (1989)). Depending on the assumptions on the specific manifold at hand it may be more convenient to use $\ell_{t|s}$ or ℓ_t^{im}. Assume that we have access to $\{\nabla \log p_{t|s} \,:\, s,t \in [0,T]\,,\ t > s\}$ or an approximation of this family, then we can use $\ell_{t|s}$, the *explicit* score function to learn $\{s_t \,:\, t \in [0,t]\}$. Using the results of Section 4.1, we highlight to methods to approximate $\ell_{t|s}$:

(a) If we have access to an approximation of $\{p_{t|0} \,:\, t \in (0,T]\,,\ t\}$ then $\ell_{t|0}$ can be used. Note that this loss is similar to the one used in the Euclidean setting, see (Song et al., 2021; Song and Ermon, 2020; Song et al., 2020; Ho et al., 2020) for instance. In the case, where the eigenvalues and the eigenfunctions of the Laplace-Beltrami operator are known then such an approximation is available, see Section 4.1. However, the quality of the approximation deteriorates when t is close to 0.

(b) If we do not have access to the eigenvalues and eigenfunctions of the Laplace-Beltrami operators then we can still derive an approximation of the $\nabla \log p_{t|s}$ for all $s \in [0,t]$ if $|t-s|$ is small enough, using Varadahn type approximations (5) and the inverse of exp[4]. In this case we use the loss functions $\ell_{t|s}$ for $|t-s|$ small enough.

[4]If $\exp^{-1}$ is not available then it can be estimated using approximated logarithmic mappings (Goto and Sato, 2021; Schiela and Ortiz, 2020) or inverse retractions (Zhu and Sato, 2020; Sato et al., 2019).

We highlight that these two methods can be used in conjunction. For instance, one can rely on the truncation techniques to estimate $\ell_{t|0}$ for large t and the Varadhan asymptotics for small t.

Last but not least, the *implicit* score loss $\{\ell_t^{\mathrm{im}}\}_{t=0}^T$ is used in cases where we do not have access to the approximations of $p_{t|s}$ for $s, t \in [0, T]$ with $t > s$. The only requirement to learn the implicit score is to be able to (approximately) sample from the forward dynamics, i.e. the Brownian motion on the Riemannian manifold. In particular, no approximation of the logarithmic derivative of the heat kernel is needed. One downside of using such an approach is that it relies on the computation of the divergence of the score s_t. The exact computation of the divergence is too costly in high dimension as it requires d Jacobian-vector calls and estimators need to be used Hutchinson (1989). Note that the loss function used in (Rozen et al., 2021) also involves computing a divergence. We summarize our different loss functions in Table 2.

Method	Loss function	Requirements
$\ell_{t\|0}$ (DSM)	$\frac{1}{2}\mathbb{E}\left[\|s(\mathbf{X}_t) - \nabla \log p_{t\|0}(\mathbf{X}_t\|\mathbf{X}_0)\|^2\right]$	▷ Sampling of $(\mathbf{X}_t, \mathbf{X}_0)$ ▷ Approximation of $\nabla \log p_{t\|0}$
$\ell_{t\|s}$ (DSM)	$\frac{1}{2}\mathbb{E}\left[\|s(\mathbf{X}_t) - \nabla \log p_{t\|0}(\mathbf{X}_t\|\mathbf{X}_s)\|^2\right]$	▷ Sampling of $(\mathbf{X}_t, \mathbf{X}_s)$ for $\|t-s\|$ small ▷ Approximation of $\nabla \log p_{t\|s}$ for $\|t-s\|$ small
ℓ_t^{im} (ISM)	$\mathbb{E}\left[\frac{1}{2}\|s(\mathbf{X}_t)\|^2 + \mathrm{div}(s)(\mathbf{X}_t)\right]$	▷ Sampling of $\mathbf{X}_t$ ▷ Approximation of $\mathrm{div}(\boldsymbol{s}_\theta)$

Table 2: Riemannian score matching losses.

Parametric family of vector fields We approximate $\{\nabla \log p_t\}_{t=0}^T$ by a family of function $\{\boldsymbol{s}_\theta\}_{\theta\in\Theta}$ where Θ is a set of parameters and for any $\theta \in \Theta$, $\boldsymbol{s}_\theta : [0, T] \to \mathcal{X}(\mathcal{M})$. In this work, we consider several parameterisations of vector fields:

- **Projected vector field**. We define $\boldsymbol{s}_\theta(t, x) = \mathrm{proj}_{T_x\mathcal{M}}(\tilde{\boldsymbol{s}}_\theta(t, x)) = P(x)\tilde{\boldsymbol{s}}_\theta(t, x)$ for any $t \in [0, T]$ and $x \in \mathcal{M}$, with $\tilde{\boldsymbol{s}}_\theta : \mathbb{R}^p \times [0, T] \to \mathbb{R}^p$ an ambient vector field and $P(x)$ the orthogonal projection over $\mathrm{T}_x\mathcal{M}$ at $x \in M$. According to Rozen et al. (2021, Lemma 2), then $\mathrm{div}(s_\theta)(x, t) = \mathrm{div}_E(s_\theta)(x, t)$ for any $x \in \mathcal{M}$, where div_E denotes the standard Euclidean divergence.
- **Divergence-free vector fields**: For any compact Lie group, any basis of the Lie algebra $\mathfrak{g}$ yields a global frame. Indeed, let $v \in \mathfrak{g}$ and define the flow $\Phi : \mathbb{R} \times \mathcal{M} \to \mathcal{M}$ given for any $t \in \mathbb{R}$ and $x \in M$ by $\Phi_t^v(x) = x\exp(tv)$. Then defining $\{E_i\}_{i=1}^d = \{\partial_t \Phi_0^{v_i}\}_{i=1}^d$, where $\{v_i\}_{i=1}^d$ is a basis of $\mathfrak{g}$, we get that $\{E_i\}_{i=1}^d$ is a left-invariant global frame. As a result, we have that for any $i \in \{1, \dots, d\}$, $\mathrm{div}(E_i) = 0$ (for the classical left invariant metric). This result simplifies the computation of $\mathrm{div}(\boldsymbol{s}_\theta)$ where $\boldsymbol{s}_\theta(t, s) = \sum_{i=1}^d \boldsymbol{s}_\theta^i(t, x)E_i(x)$ for any $t \in [0, T]$ and $x \in \mathcal{M}$ (see Falorsi and Forré, 2020).
- **Coordinates vector fields**. We define $\boldsymbol{s}_\theta(t, x) = \sum_{i=1}^d \boldsymbol{s}_\theta^i(t, x)E_i(x)$ for any $t \in [0, T]$ and $x \in \mathcal{M}$, with $\{E_i\}_{i=1}^d = \{\partial_i\varphi(x)\}_{i=1}^d$ the vector fields induced by a choice of local coordinates, where φ is a local parameterization $\varphi : \mathsf{U} \to \mathcal{M}$ and $z \in \mathsf{U} \subset \mathbb{R}^d$. Then the divergence can be computed in these local coordinates $\mathrm{div}(\boldsymbol{s}_\theta)(t, \varphi(z)) = |\det G|^{-1/2} \sum_{i=1}^d \partial_i\{|\det G|^{1/2}\boldsymbol{s}_\theta^i(t, \varphi(\cdot))\}(z)$. In the case of the sphere, one recovers the standard divergence in spherical coordinates using this formula.

VDB: I think this is needed? Maybe not

Combining this parameterization with the score-matching losses, the time-reversal formula Section 4.2 and the sampling of forward and backward processes Section 4.1, we now define our Riemannian Score-based Generative Modeling algorithm, in Algorithm 2.

EM: We do not discuss NN architectural choices for $\{s_\theta^i\}_i$ but can do for the next iteration.

4.4 Likelihood computation

Similarly to Song et al. (2021), once the score is learned we can use it in conjunction with an Ordinary Differential Equation (ODE) solver to compute the likelihood of the model. Let $\{\Phi_t\}_{t=0}^T$ be a family of vector fields. We define $(\mathbf{X}_t)_{t\in[0,T]}$ such that $\mathbf{X}_0$ has distribution p_0 (the data distribution) and satisfying $\mathrm{d}\mathbf{X}_t = \Phi_t(\mathbf{X}_t)\mathrm{d}t$. Assuming that p_0 admits a density w.r.t. p_{ref} then for any $t \in [0, T]$, the distribution of $\mathbf{X}_t$ admits a density w.r.t. p_{ref} and we denote p_t this density. We recall that $\partial_t \log p_t(\mathbf{X}_t) = \mathrm{div}(\Phi_t)(\mathbf{X}_t)$, see Mathieu and Nickel (2020, Proposition 2) for instance.

Algorithm 2 Computation of the loss

Require: $\varepsilon, T, K, p_0, \mathrm{loss}, \mathrm{thres}, s$

$\mathbf{X}_0 \sim p_0$
$t \sim U([\varepsilon, T])$ ▷ Uniform sampling between ε and T
$\mathbf{X}_t \sim \mathrm{P}_{t|0}(\mathbf{X}_0, \cdot)$ ▷ Approximate sampling using Algorithm 1
if loss = denoising **then** ▷ Denoising loss function
  if $t <$ thres **then**
    score $= -(1/t)\exp^{-1}_{\mathbf{X}_t}(\mathbf{X}_0)$ ▷ Varadhan asymptotics
  else
    score $= \sum_{j=0}^{J} \mathrm{e}^{-\lambda_j t} \nabla\phi_j(x)\phi_j(y) / \sum_{j=0}^{J} \mathrm{e}^{-\lambda_j t}\phi_j(x)\phi_j(y)$. ▷ Series truncation
  end if
  $\ell(s) = \|s(\mathbf{X}_t) - \mathrm{score}\|^2$
else ▷ Implicit loss function
  $\ell(s) = (1/2)\|s(\mathbf{X}_t)\|^2 + \mathrm{div}(s)(\mathbf{X}_t)$
end if
return $\ell(s)$

Recall that we consider a Brownian motion on the manifold as a forward process $(\mathbf{B}_t^{\mathcal{M}})_{t\in[0,T]}$ with $\{p_t\}_{t=0}^T$ the associated family of densities. Thus we have that for any $t \in [0,T]$ and $x \in \mathcal{M}$

$$\partial_t p_t(x) = \tfrac{1}{2}\Delta p_t(x) = \mathrm{div}\left(\tfrac{1}{2} p_t \nabla \log p_t\right)(x).$$

Hence, we can define $(\mathbf{X}_t)_{t\in[0,T]}$ satisfying $\mathrm{d}\mathbf{X}_t = \frac{1}{2}\nabla \log p_t(\mathbf{X}_t)\mathrm{d}t$ such that $\mathbf{X}_0$ has distribution p_0. Defining $(\hat{\mathbf{X}}_t)_{t\in[0,T]} = (\mathbf{X}_{T-t})_{t\in[0,T]}$, it follows that $\hat{\mathbf{X}}_0$ has distribution $\mathcal{L}(\mathbf{X}_T)$ and satisfies

$$\mathrm{d}\hat{\mathbf{X}}_t = -\tfrac{1}{2}\nabla \log p_{T-t}(\hat{\mathbf{X}}_t)\mathrm{d}t. \tag{7}$$

Finally, we introduce $(\mathbf{Y}_t)_{t\in[0,T]}$ satisfying (7) but such that $\mathbf{Y}_0 \sim p_{\mathrm{ref}}$. Note that if $T \geq 0$ is large then the two processes $(\mathbf{Y}_t)_{t\in[0,T]}$ and $(\hat{\mathbf{X}}_t)_{t\in[0,T]}$ are close since $\mathcal{L}(\mathbf{X}_T)$ is close to p_{ref}. Therefore, using the score network and a manifold ODE solver (as in Mathieu and Nickel, 2020), we are able to approximately solve the following ODE

$$\partial_t \log q_t(\mathbf{Y}_t) = -\tfrac{1}{2}\mathrm{div}(\boldsymbol{s}_\theta(t,\cdot))(\mathbf{Y}_t),$$

with q_t the density of $\mathbf{Y}_t$ w.r.t. p_{ref} and $\log q_0(\mathbf{Y}_0) = 0$. The likelihood approximation of the model is then given by $\log q_T(\mathbf{Y}_T)$. In Appendix S3, we highlight that this likelihood computation is slightly different from the one obtained using the SDE.

5 Related work

In what follows, we discuss previous work on parametrizing family of distributions for manifold-valued data. Note that in this work, the manifold structure is considered to be prescribed. In contrast, another line of work has been focusing on jointly learning the manifold structure and a generative model (Brehmer and Cranmer, 2020; Kalatzis et al., 2021; Caterini et al., 2021).

Parametric family of distributions. Defining flexible easy-to-sample distributions on manifolds is not a trivial task. The various parametric families of distributions that have been proposed can broadly be categorised into three main approaches (Navarro et al., 2017): wrapping, projecting and conditioning. Wrapped distributions consider a parametric distribution on $\mathbb{R}^n$ that is pushed-forward along an invertible map $\psi : \mathbb{R}^n \to \mathcal{M}$. A canonical example is the wrapped normal distribution on $\mathbb{S}^1$ (Collett and Lewis, 1981). Another example has been proposed by Mathieu et al. (2019); Nagano et al. (2019) on the hyperbolic space with the exponential map . Given a Euclidean submanifold $\mathcal{M} \subset \mathbb{R}^n$ and a distribution $p_{\mathrm{amb}} \in \mathcal{P}(\mathbb{R}^n)$, marginalizing out p_{amb} along the normal bundle induces a distribution on $\mathcal{M}$. Samples are obtained by first sampling p_{amb} and then applying an orthogonal projection on these samples. Finally, the conditioning method consists into considering the unormalized density defined by the restriction of an ambient density p_{amb} with $\mathcal{M}$. Such distributions encompass the von Mises-Fisher distribution (Fisher, 1953) and the Kent distribution (Kent, 1982). These distributions are usually unimodal and considering mixtures of thereof is key to increase flexibility (Peel et al., 2001; Mardia et al., 2008).

VDB: this is the same thing as push forward of Euclidean NF?

Push-forward of Euclidean normalizing flows. More recently, approaches leveraging the flexibility of normalizing flows (Papamakarios et al., 2019) have been proposed. Following the wrapping method described above, these methods parametrize a normalizing flow in the Euclidean space $\mathbb{R}^n$ that is pushed-forward along an invertible map $\psi : \mathbb{R}^n \to \mathcal{M}$. However, to globally represent the manifold, the map ψ needs to be a homeomorphism, which can only happen if $\mathcal{M}$ is topologically equivalent to $\mathbb{R}^n$, hence limiting the scope of that approach. One natural choice for this map if the exponential map $\exp_x : \mathrm{T}_x\mathcal{M} \cong \mathbb{R}^d$. This approach has been taken, for instance, by Falorsi et al. (2019) and Bose et al. (2020), respectively parametrizing distributions on Lie groups and hyperbolic space.

Neural ODE on manifolds. To avoid artifacts or numerical instabilities due to the manifold embedding, another line of work uses tools from Riemannian geometry to define flows directly on the manifold of interest (Falorsi and Forré, 2020; Mathieu and Nickel, 2020; Falorsi, 2021). Since these methods do not require a specific embedding mapping, they are referred as *Riemannian*. They extend continuous normalizing flows (CNFs) (Grathwohl et al., 2019) to the manifold setting, by implicity parametrizing flows as solutions of Ordinary Differential Equations (ODEs). As such, the parametric flow is a *continuous* function of time. This approach has recently been extended by Rozen et al. (2021) introducing Moser flows, whose main appeal being that it circumvents the need to solve an ODE in the training process.

Optimal transport on manifolds. Another line of work has focused on developing flows on manifolds using tools from optimal transport. Sei (2013) introduced a flow that is given by $f_\theta : x \mapsto \exp_x(\nabla\psi_\theta^c)$ with ψ_θ^c a c-convex function and $c = d_{\mathcal{M}}^2$, where $d_{\mathcal{M}}$ is the geodesic distance. This approach is motivated by the fact that the optimal transport map takes such an expression (Ambrosio, 2003). These methods operate directly on the manifold, similarly to CNFs, yet in contrast they are *discrete* in time. The benefits of this approach depend on the specific choice of parametric family of c-convex functions (Rezende and Racanière, 2021; Cohen et al., 2021), trading-off expressively with scalability. In the case of tori and spheres, Rezende et al. (2020) introduced *discrete* Riemannian flows based on Möbius transformations and spherical splines.

6 Experiments

We evaluate the model on a collection of datasets, each containing an empirical distribution of occurrences of earth and climate science events on the surface of the earth. These events are: volcanic eruptions (NGDC/WDS), earthquakes (NGDC/WDS), floods (Brakenridge, 2017) and wild fires (EOSDIS, 2020). In each case the earth is approximated as a perfect sphere. We compare to previous baseline methods: Riemannian Continuous Normalizing Flows (Mathieu and Nickel, 2020), Moser Flows (Rozen et al., 2021) and a mixture of Kent distributions (Peel et al., 2001). The mixture of Kent distributions is optimised using an EM algorithm and the optimal number of components is selected on a validation set. Additionally, we consider another score-based generative model: a standard SBGM on the 2D place followed by the inverse stereographic projection which induces a density on the sphere (Gemici et al., 2016). More experimental details can be found in Appendix S8. We observe from Table 3, that the RSBGM model outperforms all other methods in density estimation, in particular by a large margin on the volcanic eruptions dataset.

	Volcano	**Earthquake**	**Flood**	**Fire**
Mixture of Kent	$-0.95_{\pm 0.14}$	$0.14_{\pm 0.13}$	$0.73_{\pm 0.07}$	$-1.18_{\pm 0.06}$
Riemannian CNF	$-0.97_{\pm 0.15}$	$0.19_{\pm 0.04}$	$0.90_{\pm 0.03}$	$-0.66_{\pm 0.05}$
Moser Flow	$-2.02_{\pm 0.42}$	$-0.09_{\pm 0.02}$	$0.62_{\pm 0.04}$	$-1.03_{\pm 0.03}$
Stereographic Score-Based	$-4.37_{\pm ???}$	$-0.05_{\pm ???}$	$1.32_{\pm ???}$	$0.11_{\pm ???}$
Riemannian Score-Based	$\mathbf{-5.56}_{\pm 0.26}$	$\mathbf{-0.21}_{\pm 0.03}$	$\mathbf{0.52}_{\pm 0.02}$	$\mathbf{-1.24}_{\pm 0.07}$
Dataset size	827	6120	4875	12809

Table 3: Negative log-likelihood scores for each method on the earth and climate science datasets. Bold indicates statistically significant best method. Means and standard deviations are computed over 5 different runs.

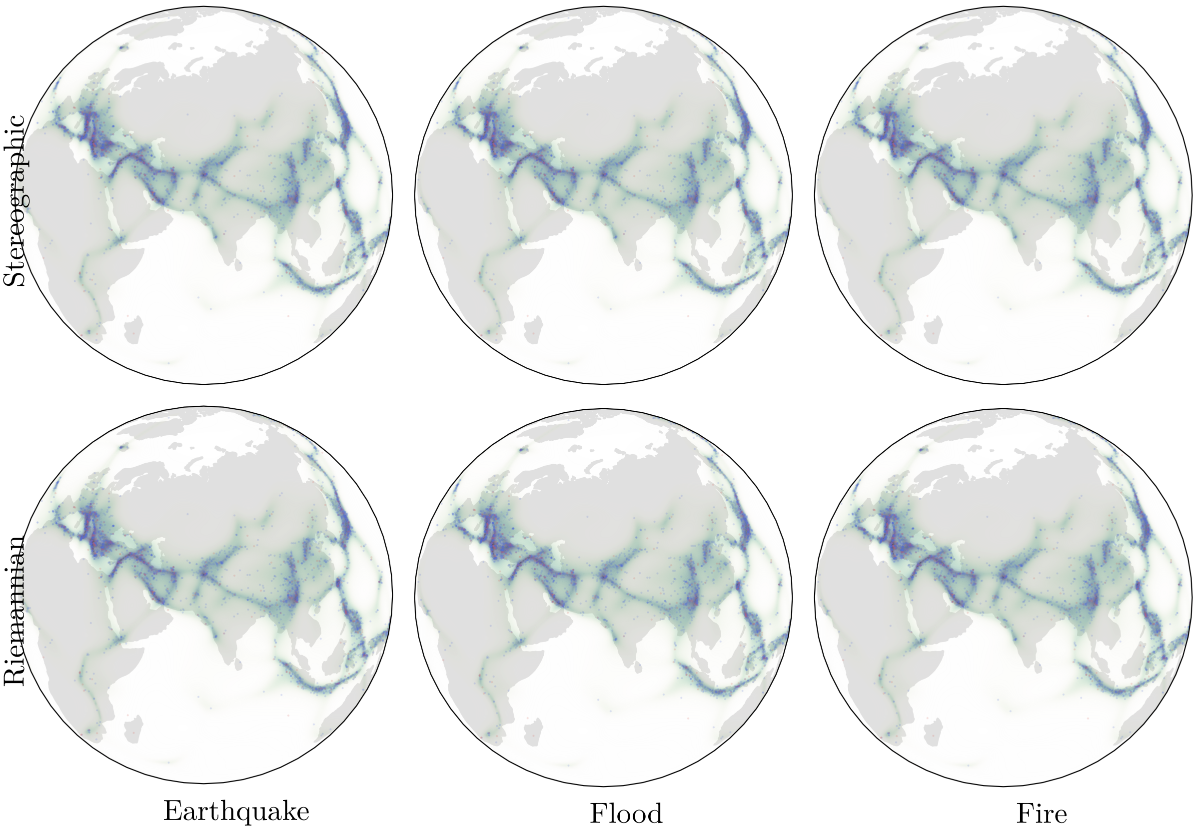

Figure 1: Trained score-based generative models on earth sciences data. The learned density is colored green-blue. Blue and red dots represent training and testing datapoints, respectively.

7 Discussion and limitations

In this paper we introduced Riemannian Score-Based Generative Models (RSGMs), a class of deep generative models that represent target densities supported on compact manifolds, as the reverse diffusion process of a Brownian motion. The main benefits of our method stems from its scalability to high dimensions, its applicability to a broad class of manifolds due to the diversity of available loss functions and its capacity to model complex datasets. We proved that RSGMs are universal generative models for densities supported on compact manifolds. Empirically, we demonstrated that our method outperforms previous work on density estimation tasks with spherical geoscience datasets.

VDB: universal ?

One current limitation—similarly to other score-based generative models—is the requirement of samples from the targeted distribution, as such models cannot directly fit an unnormalized density. An important future work direction, and a current limitation, is the manifold compactness assumption. Several important manifolds do not fit into this category, such as the special linear group, symmetric positive definite matrices or the hyperbolic space which underlies special relativity (Ungar, 2005). Another extension of interest is conditional sampling. By amortizing SGMs with respect to an observation it is possible to approximately sample from a given posterior distribution (see for instance Kawar et al., 2021a;b; Lee et al., 2021; Sinha et al., 2021; Batzolis et al., 2021; Chung et al., 2021).

EM: Should write a paragraph in the app on the conditional extension

References

Ambrosio, L. (2003). Optimal transport maps in Monge-Kantorovich problem. *arXiv preprint arXiv:0304389v1*.

Atkinson, K. and Han, W. (2012). *Spherical Harmonics and Approximations on the Unit Sphere: An Introduction*, volume 2044. Springer Science & Business Media.

Bao, F., Li, C., Zhu, J., and Zhang, B. (2022). Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. *arXiv preprint arXiv:2201.06503*.

Batzolis, G., Stanczuk, J., Schönlieb, C.-B., and Etmann, C. (2021). Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*.

Bismut, J.-M. (1984). Large deviations and the Malliavin calculus. *Birkhauser Prog. Math.*, 45.

Boomsma, W., Mardia, K. V., Taylor, C. C., Ferkinghoff-Borg, J., Krogh, A., and Hamelryck, T. (2008). A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences*, 105(26):8932–8937.

Bose, J., Smofsky, A., Liao, R., Panangaden, P., and Hamilton, W. (2020). Latent variable modelling with hyperbolic normalizing flows. In *International Conference on Machine Learning*, pages 1045–1055. PMLR.

Brakenridge, G. (2017). Global active archive of large flood events. http://floodobservatory.colorado.edu/Archives/index.html.

Brehmer, J. and Cranmer, K. (2020). Flows for simultaneous manifold learning and density estimation. *arXiv preprint arXiv:2003.13913*.

Caterini, A. L., Loaiza-Ganem, G., Pleiss, G., and Cunningham, J. P. (2021). Rectangular flows for manifold learning. *arXiv preprint arXiv:2106.01413*.

Cattiaux, P., Conforti, G., Gentil, I., and Léonard, C. (2021). Time reversal of diffusion processes under a finite entropy condition. *arXiv preprint arXiv:2104.07708*.

Chen, T., Liu, G.-H., and Theodorou, E. A. (2021a). Likelihood training of Schrödinger bridge using forward-backward sdes theory. *arXiv preprint arXiv:2110.11291*.

Chen, X., Li, X. M., and Wu, B. (2021b). Logarithmic heat kernels: estimates without curvature restrictions. *arXiv preprint arXiv:2106.02746*.

Chen, Y., Georgiou, T., and Pavon, M. (2016). Entropic and displacement interpolation: a computational approach using the Hilbert metric. *SIAM Journal on Applied Mathematics*, 76(6):2375–2396.

Chung, H., Sim, B., and Ye, J. C. (2021). Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. *arXiv preprint arXiv:2112.05146*.

Cohen, S., Amos, B., and Lipman, Y. (2021). Riemannian convex potential maps. *arXiv preprint arXiv:2106.10272*.

Collett, D. and Lewis, T. (1981). Discriminating Between the Von Mises and Wrapped Normal Distributions. *Australian Journal of Statistics*, 23(1):73–79.

De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. (2021). Diffusion Schrödinger bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems*.

Dhariwal, P. and Nichol, A. (2021). Diffusion models beat GAN on image synthesis. *arXiv preprint arXiv:2105.05233*.

Dormand, R. J. and Prince, J. P. (1980). A family of embedded Runge-Kutta formulae. *Journal of Computational and Applied Mathematics*, pages 19–26.

Durkan, C. and Song, Y. (2021). On maximum likelihood training of score-based generative models. *arXiv preprint arXiv:2101.09258*.

EOSDIS (2020). Land, atmosphere near real-time capability for eos (lance) system operated by nasa's earth science data and information system (esdis). https://earthdata.nasa.gov/earth-observation-data/near-real-time/firms/active-fire-data.

Falorsi, L. (2021). Continuous normalizing flows on manifolds. *arXiv:2104.14959*.

Falorsi, L., de Haan, P., Davidson, T. R., and Forré, P. (2019). Reparameterizing distributions on lie groups. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3244–3253. PMLR.

Falorsi, L. and Forré, P. (2020). Neural ordinary differential equations on manifolds. *arXiv preprint arXiv:2006.06663*.

Federer, H. (2014). *Geometric Measure Theory*. Springer.

Feiten, W., Lang, M., and Hirche, S. (2013). Rigid motion estimation using mixtures of projected gaussians. In *Proceedings of the 16th International Conference on Information Fusion*, pages 1465–1472. IEEE.

Fisher, R. A. (1953). Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 217(1130):295–305.

García-Zelada, D. and Huguet, B. (2021). Brenier–Schrödinger problem on compact manifolds with boundary. *Stochastic Analysis and Applications*, pages 1–29.

Gemici, M. C., Rezende, D., and Mohamed, S. (2016). Normalizing flows on Riemannian manifolds. *arXiv preprint arXiv:1611.02304*.

Goto, J. and Sato, H. (2021). Approximated logarithmic maps on Riemannian manifolds and their applications. *JSIAM Letters*, 13:17–20.

Grathwohl, W., Chen, R. T. Q., Bettencourt, J., and Duvenaud, D. (2019). Scalable reversible generative models with free-form continuous dynamics. In *International Conference on Learning Representations*.

Gunther, M. (1991). Isometric embeddings of Riemannian manifolds, Kyoto, 1990. In *Proc. Intern. Congr. Math.*, pages 1137–1143. Math. Soc. Japan.

Hamelryck, T., Kent, J. T., and Krogh, A. (2006). Sampling realistic protein conformations using local structural bias. *PLoS Computational Biology*, 2(9):e131.

Haussmann, U. G. and Pardoux, E. (1986). Time reversal of diffusions. *The Annals of Probability*, 14(4):1188–1205.

He, Y. (2013). A lower bound for the first eigenvalue in the laplacian operator on compact riemannian manifolds. *Journal of Geometry and Physics*, 71:73–84.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*.

Hsu, E. P. (2002). *Stochastic Analysis on Manifolds*. Number 38. American Mathematical Society.

Huang, C.-W., Lim, J. H., and Courville, A. (2021). A variational perspective on diffusion-based generative models and score matching. *arXiv preprint arXiv:2106.02808*.

Hutchinson, M. F. (1989). A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076.

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

Ikeda, N. and Watanabe, S. (1989). *Stochastic Differential Equations and Diffusion Processes*, volume 24 of *North-Holland Mathematical Library*. North-Holland Publishing Co., Amsterdam; Kodansha, Ltd., Tokyo, second edition.

Jolicoeur-Martineau, A., Li, K., Piché-Taillefer, R., Kachman, T., and Mitliagkas, I. (2021a). Gotta go fast when generating data with score-based models. *arXiv preprint arXiv:2105.14080*.

Jolicoeur-Martineau, A., Piché-Taillefer, R., Tachet des Combes, R., and Mitliagkas, I. (2021b). Adversarial score matching and improved sampling for image generation. *International Conference on Learning Representations*.

Jones, P. W., Maggioni, M., and Schul, R. (2008). Manifold Parametrizations by Eigenfunctions of the Laplacian and Heat Kernels. *Proceedings of the National Academy of Sciences of the United States of America*, 105(6):1803–1808.

Jørgensen, E. (1975). The central limit problem for geodesic random walks. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32(1-2):1–64.

Kalatzis, D., Ye, J. Z., Wohlert, J., and Hauberg, S. (2021). Multi-chart flows. *arXiv preprint arXiv:2106.03500*.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., and Kumar, V. (2018). Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, 31(8):1544–1554.

Kawar, B., Vaksman, G., and Elad, M. (2021a). Snips: Solving noisy inverse problems stochastically. *arXiv preprint arXiv:2105.14951*.

Kawar, B., Vaksman, G., and Elad, M. (2021b). Stochastic image denoising by sampling from the posterior distribution. *arXiv preprint arXiv:2101.09552*.

Kent, J. T. (1982). The Fisher-Bingham distribution on the sphere. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(1):71–80.

Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*.

Klimovskaia, A., Lopez-Paz, D., Bottou, L., and Nickel, M. (2020). Poincaré maps for analyzing complex hierarchies in single-cell data. *Nature communications*, 11(1):1–9.

Kloeden, P. and Platen, E. (2011). *Numerical Solution of Stochastic Differential Equations*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg.

Kobayashi, S. and Nomizu, K. (1963). *Foundations of Differential Geometry*, volume 1. New York, London.

Kolár, I., Michor, P. W., and Slovák, J. (2013). *Natural Operations in Differential Geometry.* Springer Science & Business Media.

Kurtz, T. G., Pardoux, É., and Protter, P. (1995). Stratonovich stochastic differential equations driven by general semimartingales. In *Annales de l'IHP Probabilités et statistiques*, volume 31, pages 351–377.

Lee, J. (2010). *Introduction to Topological Manifolds*, volume 202. Springer Science & Business Media.

Lee, J. M. (2006). *Riemannian Manifolds: An Introduction to Curvature*, volume 176. Springer Science & Business Media.

Lee, J. M. (2013). Smooth manifolds. In *Introduction to Smooth Manifolds*, pages 1–31. Springer.

Lee, J. M. (2018). *Introduction to Riemannian manifolds.* Springer.

Lee, S.-g., Kim, H., Shin, C., Tan, X., Liu, C., Meng, Q., Qin, T., Chen, W., Yoon, S., and Liu, T.-Y. (2021). Priorgrad: Improving conditional denoising diffusion models with data-driven adaptive prior. *arXiv preprint arXiv:2106.06406.*

Leobacher, G. and Steinicke, A. (2021). Existence, uniqueness and regularity of the projection onto differentiable manifolds. *Annals of Global Analysis and Geometry*, 60(3):559–587.

Léonard, C. (2012a). From the Schrödinger problem to the Monge–Kantorovich problem. *Journal of Functional Analysis*, 262(4):1879–1920.

Léonard, C. (2012b). Girsanov theory under a finite entropy condition. In *Séminaire de Probabilités XLIV*, pages 429–465. Springer.

Léonard, C., Rœlly, S., Zambrini, J.-C., et al. (2014). Reciprocal processes: a measure-theoretical point of view. *Probability Surveys*, 11:237–269.

Li, P. (1986). Large time behavior of the heat equation on complete manifolds with non-negative ricci curvature. *Annals of Mathematics*, 124(1):1–21.

Lou, A., Lim, D., Katsman, I., Huang, L., Jiang, Q., Lim, S.-N., and De Sa, C. (2020). Neural manifold ordinary differential equations. *arXiv preprint arXiv:2006.10254.*

Lui, Y. M. (2012). Advances in matrix manifolds for computer vision. *Image and Vision Computing*, 30(6-7):380–388.

Mardia, K. V., Hughes, G., Taylor, C. C., and Singh, H. (2008). A multivariate von Mises distribution with applications to bioinformatics. *Canadian Journal of Statistics*, 36(1):99–109.

Mardia, K. V., Taylor, C. C., and Subramaniam, G. K. (2007). Protein bioinformatics and mixtures of bivariate von mises distributions for angular data. *Biometrics*, 63(2):505–512.

Mathieu, E., Lan, C. L., Maddison, C. J., Tomioka, R., and Teh, Y. W. (2019). Continuous hierarchical representations with poincar\'e variational auto-encoders. *arXiv preprint arXiv:1901.06033.*

Mathieu, E. and Nickel, M. (2020). Riemannian continuous normalizing flows. *arXiv preprint arXiv:2006.10605.*

Nagano, Y., Yamaguchi, S., Fujita, Y., and Koyama, M. (2019). A wrapped normal distribution on hyperbolic space for gradient-based learning. In *International Conference on Machine Learning*, pages 4693–4702. PMLR.

Navarro, A. K., Frellsen, J., and Turner, R. E. (2017). The multivariate generalised von mises distribution: inference and applications. In *Thirty-First AAAI Conference on Artificial Intelligence.*

(NGDC/WDS), N. G. D. C. . W. D. S. (2022a). Ncei/wds global significant earthquake database. https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ngdc.mgg.hazards:G012153.

(NGDC/WDS), N. G. D. C. . W. D. S. (2022b). Ncei/wds global significant volcanic eruptions database. https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ngdc.mgg.hazards:G10147.

Nichol, A. and Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. *arXiv preprint arXiv:2102.09672.*

Nutz, M. and Wiesel, J. (2022). Stability of Schrödinger potentials and convergence of Sinkhorn's algorithm. *arXiv preprint arXiv:2201.10059.*

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2019). Normalizing flows for probabilistic modeling and inference. *arXiv preprint arXiv:1912.02762.*

Peel, D., Whiten, W. J., and McLachlan, G. J. (2001). Fitting mixtures of kent distributions to aid in joint set identification. *Journal of the American Statistical Association*, 96(453):56–63.

Peyré, G. and Cuturi, M. (2019). Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Revuz, D. and Yor, M. (1999). *Continuous Martingales and Brownian Motion*, volume 293 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, third edition.

Rezende, D. J., Papamakarios, G., Racanière, S., Albergo, M. S., Kanwar, G., Shanahan, P. E., and Cranmer, K. (2020). Normalizing flows on tori and spheres. *arXiv:2002.02428*.

Rezende, D. J. and Racanière, S. (2021). Implicit Riemannian concave potential maps. *arXiv:2110.01288*.

Roy, D. M., Kemp, C., Mansinghka, V., and B Tenenbaum, J. (2007). Learning annotated hierarchies from relational data.

Rozen, N., Grover, A., Nickel, M., and Lipman, Y. (2021). Moser flow: Divergence-based generative modeling on manifolds. *Advances in Neural Information Processing Systems*, 34.

Saloff-Coste, L. (1994). Precise estimates on the rate at which certain diffusions tend to equilibrium. *Mathematische Zeitschrift*, 217(1):641–677.

Sato, H., Kasai, H., and Mishra, B. (2019). Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport. *SIAM Journal on Optimization*, 29(2):1444–1472.

Schiela, A. and Ortiz, J. (2020). An SQP method for equality constrained optimization on manifolds. *arXiv preprint arXiv:2005.06844*.

Schrödinger, E. (1932). Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. *Annales de l'Institut Henri Poincaré*, 2(4):269–310.

Sei, T. (2013). A Jacobian inequality for gradient maps on the sphere and its application to directional statistics. *Communications in Statistics-Theory and Methods*, 42(14):2525–2542.

Senanayake, R. and Ramos, F. (2018). Directional grid maps: modeling multimodal angular uncertainty in dynamic environments. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3241–3248. IEEE.

Shapovalov, M. V. and Dunbrack Jr, R. L. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, 19(6):844–858.

Sinha, A., Song, J., Meng, C., and Ermon, S. (2021). D2c: Diffusion-denoising models for few-shot conditional generation. *arXiv preprint arXiv:2106.06819*.

Sinkhorn, R. (1967). Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405.

Song, J., Meng, C., and Ermon, S. (2020). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*.

Song, Y. and Ermon, S. (2020). Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems*.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021). Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

Steyvers, M. and Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, 29(1):41–78.

Ungar, A. (2005). Einstein's special relativity: Unleashing the power of its hyperbolic geometry. *Computers & Mathematics with Applications*, 49(2):187–221.

Urakawa, H. (2006). Convergence rates to equilibrium of the heat kernels on compact Riemannian manifolds. *Indiana University mathematics journal*, pages 259–288.

Vargas, F., Thodoroff, P., Lawrence, N. D., and Lamacraft, A. (2021). Solving Schrödinger bridges via maximum likelihood. *arXiv preprint arXiv:2106.02081*.

Vincent, P. (2011). A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674.

Watson, D., Ho, J., Norouzi, M., and Chan, W. (2021). Learning to efficiently sample from diffusion probabilistic models. *arXiv preprint arXiv:2106.03802*.

Zhu, X. and Sato, H. (2020). Riemannian conjugate gradient methods with inverse retraction. *Computational Optimization and Applications*, 77(3):779–810.

Supplementary to: Riemannian Score-Based Generative Modeling

S1 Organization of the supplementary

In this supplementary we gather the proof of Theorem 5 as well as additional derivations on score-based generative models and Riemannian manifolds. In Appendix S2, we recall basics on stochastic Riemannian geometry following Hsu (2002). In Appendix S3, we highlight differences between ODE and SDE models for likelihood computation. In Appendix S4, we recall some basic facts about eigenvalues and eigenfunctions of the Laplace-Beltrami operator on the d-dimensional sphere and torus. In Appendix S5, we present the extension of the time-reversal formula to manifold and prove Theorem 5.

S2 Preliminaries on stochastic Riemannian geometry

In this section, we recall some basic facts on Riemannian geometry and stochastic Riemannian geometry. We follow Hsu (2002); Lee (2018; 2006) and refer to Lee (2010; 2013) for a general introduction to topological and smooth manifolds. Throughout this section $\mathcal{M}$ is a d-dimensional smooth manifold, $\mathrm{T}\mathcal{M}$ its tangent bundle and $\mathrm{T}^\star\mathcal{M}$ it cotangent bundle. We denote $\mathrm{C}^\infty(\mathcal{M})$ the set of real-valued smooth functions on $\mathcal{M}$ and $\mathcal{X}(\mathcal{M})$ the set of vector fields on $\mathcal{M}$.

S2.1 Tensor field, metric, connection and transport

Tensor field and Riemannian metric For a vector space V let $\mathrm{T}^{k,\ell}(V) = V^{\otimes k} \otimes (V^\star)^{\otimes \ell}$ with $k, \ell \in \mathbb{N}$. For any $k, \ell \in \mathbb{N}$ we define the space of (k,ℓ)-tensors as $\mathrm{T}^{k,\ell}\mathcal{M} = \sqcup_{p\in\mathcal{M}} \mathrm{T}^{k,\ell}(\mathrm{T}_p\mathcal{M})$. Note that $\Gamma(\mathcal{M}, \mathrm{T}^{0,0}\mathcal{M}) = \mathrm{C}^\infty(\mathcal{M})$, $\mathcal{X}(\mathcal{M}) = \Gamma(\mathcal{M}, \mathrm{T}^{1,0}\mathcal{M})$ and that the space of 1-form on $\mathcal{M}$ is given by $\Gamma(\mathcal{M}, \mathrm{T}^{0,1}\mathcal{M})$, where $\Gamma(\mathcal{M}, V(\mathcal{M}))$ is a section of a vector bundle $V(\mathcal{M})$ (see Lee, 2013, Chapter 10). For any $k \in \mathbb{N}$, we denote $\mathrm{T}^{|k|}\mathcal{M} = \sqcup_{j=0}^k \mathrm{T}^{j,k-j}\mathcal{M}$. $\mathcal{M}$ is said to be a Riemannian manifold if there exists $g \in \Gamma(\mathcal{M}, \mathrm{T}^{0,2}\mathcal{M})$ such that for any $x \in \mathcal{M}$, $g(x)$ is positive definite. g is called the Riemannian metric of $\mathcal{M}$. Every smooth manifold can be equipped with a Riemannian metric (see Lee, 2018, Proposition 2.4). In local coordinates we define $G = \{g_{i,j}\}_{1\le i,j\le d} = \{g(X_i, X_j)\}_{1\le i,j\le d}$, where $\{X_i\}_{i=1}^d$ is a basis of the tangent space. In what follows we consider that $\mathcal{M}$ is equipped with a metric g and for any $X, Y \in \mathcal{X}(\mathcal{M})$ we denote $\langle X, Y\rangle_\mathcal{M} = g(X, Y)$.

Connection A connection ∇ is a mapping which allows one to differentiate vector fields w.r.t other vector fields. ∇ is a linear map $\nabla : \mathcal{X}(\mathcal{M}) \times \mathcal{X}(\mathcal{M}) \to \mathcal{X}(\mathcal{M})$. In addition, we assume that i) for any $f \in \mathrm{C}^\infty(\mathcal{M})$, $X, Y \in \mathcal{X}(\mathcal{M})$, $\nabla_{fX}(Y) = f\nabla_X Y$, ii) for any $f \in \mathrm{C}^\infty(\mathcal{M})$, $X, Y \in \mathcal{X}(\mathcal{M})$, $\nabla_X(fY) = f\nabla_X Y + X(f)Y$. Given a system of local coordinates, the Christoffel symbols $\{\Gamma_{i,j}^k\}_{1\le i,j,k\le d}$ are given for any $i, j \in \{1, \dots, d\}$ by $\nabla_{X_i} X_j = \sum_{k=1}^d \Gamma_{i,j}^k X_k$. We also define the Levi-Civita connection ∇ by considering the additional two conditions: i) ∇ is torsion-free, *i.e.* for any $X, Y \in \mathcal{X}(\mathcal{M})$ we have $\nabla_X Y - \nabla_Y X = [X, Y]$, where $[X, Y]$ is the Lie bracket between X and Y, ii) ∇ is compatible with the metric g, *i.e.* for any $X, Y, Z \in \mathcal{X}(\mathcal{M})$, $X(\langle Y, Z\rangle_\mathcal{M}) = \langle \nabla_X Y, Z\rangle_\mathcal{M} + \langle Y, \nabla_X Z\rangle_\mathcal{M}$. We recall that the Levi-Civita connection is uniquely defined since for any $X, Y, Z \in \mathcal{X}(\mathcal{M})$ we have

$$2\langle \nabla_X Y, Z\rangle_\mathcal{M} = X(\langle Y, Z\rangle_\mathcal{M}) + Y(\langle Z, X\rangle_\mathcal{M}) - Z(\langle X, Y\rangle_\mathcal{M}) + \langle [X, Y], Z\rangle_\mathcal{M} - \langle [Z, X], Y\rangle_\mathcal{M} - \langle [Y, Z], X\rangle_\mathcal{M}.$$

In this case, we have that the Christoffel symbols are given for any $i, j, k \in \{1, \dots, d\}$ by

$$\Gamma_{i,j}^k = (1/2) \textstyle\sum_{m=1}^d g^{km}(\partial_j g_{m,i} + \partial_i g_{m,j} - \partial_m g_{i,j}),$$

where $\{g^{i,j}\}_{1\le i,j\le d} = G^{-1}$. Note that if $\mathcal{M}$ is Euclidean then for any $i, j, k \in \{1, \dots, d\}$, $\Gamma_{i,j}^k = 0$. We also extend the connection so that for any $X \in \mathcal{X}(\mathcal{M})$ and $f \in \mathrm{C}^\infty(M)$ we have $\nabla_X f = X(f)$. In particular, we have that $\nabla_X f \in \mathrm{C}^\infty(\mathcal{M})$. In addition, we extend the connection such that for any $\alpha \in \Gamma(\mathcal{M}, \mathrm{T}^{0,1}\mathcal{M})$, $X, Y \in \mathcal{X}(\mathcal{M})$ we have $\nabla_X \alpha(Y) = \alpha(\nabla_X Y) - X(\alpha(Y))$. In particular, we have that $\nabla_X \alpha \in \Gamma(\mathcal{M}, \mathrm{T}^{1,0}\mathcal{M})$. Note that for any $X \in \mathcal{X}(\mathcal{M})$ and $\alpha, \beta \in \mathrm{T}^{|1|}\mathcal{M}$ we have $\nabla_X(\alpha \otimes \beta) = \nabla_X \alpha \otimes \beta + \alpha \otimes \nabla_X \beta$. Similarly, we can define recursively $\nabla_X \alpha$ for any $\alpha \in \Gamma(\mathcal{M}, \mathrm{T}^{k,\ell}\mathcal{M})$ with $k, \ell \in \mathbb{N}$. Such an extension is called a covariant derivative.

Parallel transport, geodesics and exponential mapping Given a connection, we can define the notion of parallel transport, which transports vector fields along a curve. Let $\gamma : [0,1] \to \mathcal{M}$ be a smooth curve. We define the covariant derivative along the curve γ by $D_{\dot{\gamma}} : \mathcal{X}(\gamma) \to \mathcal{X}(\gamma)$ similarly to the connection, where $\mathcal{X}(\gamma) = \Gamma(\gamma([0,1]), \mathrm{T}\mathcal{M})$. In particular if $\dot{\gamma}$ and $X \in \mathcal{X}(\gamma)$ can be extended to $\mathcal{X}(\mathcal{M})$ then we define $D_{\dot{\gamma}}(X) = \nabla_{\dot{\gamma}} X \in \mathcal{X}(\mathcal{M})$. In what follows, we denote $D = \nabla$ for simplicity. We say that $X \in \mathcal{X}(\gamma)$ is parallel to γ if for any $t \in [0,1]$, $\nabla_{\dot{\gamma}} X(t) = 0$. In local coordinates, let $X \in \mathcal{X}(\gamma)$ be given for any $t \in [0,1]$ by $X = \sum_{i=1}^d a_i(t) E_i(t)$ (assuming that $\gamma([0,1])$ is entirely contained in a local chart), then we have that for any $t \in [0,1]$ and $k \in \{1, \dots, d\}$

$$\dot{a}_k(t) + \sum_{i,j=1}^d \Gamma_{i,j}^k(x(t)) \dot{x}_i(t) a_j(t) = 0. \tag{S1}$$

A curve γ on $\mathcal{M}$ is said to be a geodesics if $\dot{\gamma}$ is parallel to γ. Using Equation (S1) we get that

$$\ddot{x}_k(t) + \sum_{i,j=1}^d \Gamma_{i,j}^k(x(t)) \dot{x}_i(t) \dot{x}_j(t) = 0.$$

For more details on geodesics and parallel transport, we refer to Lee (2018, Chapter 4). Parallel transport will be key to define the frame bundle and the orthonormal frame bundle in Appendix S2.3. In addition, we have that parallel transport provides a linear isomorphism between tangent spaces. Indeed, let $v \in \mathrm{T}_x\mathcal{M}$ and $\gamma : [0,1] \to \mathcal{M}$ with $\gamma(0) = x$ a smooth curve. Then, there exists a unique vector field $X^v \in \mathcal{X}(\gamma)$ such that $X^v(x) = v$ and X^v is parallel to γ. For any $t \in [0,1]$, we denote $\Gamma_0^t : \mathrm{T}_x\mathcal{M} \to \mathrm{T}_{\gamma(t)}\mathcal{M}$ the linear isomorphism such that $\Gamma_0^t(v) = X^v(\gamma(t))$.

For any $x \in \mathcal{M}$ and $v \in \mathrm{T}_x\mathcal{M}$ we denote $\gamma^{x,v} : [0, \varepsilon^{x,v}]$ the geodesics (defined on the maximal interval $[0, \varepsilon^{x,v}]$) on $\mathcal{M}$ such that $\gamma(0) = x$ and $\dot{\gamma}(0) = v$. We denote $\mathsf{U}^x = \{v \in \mathrm{T}_x\mathcal{M} : \varepsilon^{x,v} \geq 1\}$. Note that $0 \in \mathsf{U}^x$. For any $x \in \mathcal{M}$, we define the exponential mapping $\exp_x : \mathsf{U}^x \to \mathcal{M}$ such that for any $v \in \mathsf{U}^x$, $\exp_x(v) = \gamma^{x,v}(1)$. If for any $x \in \mathcal{M}$, $\mathsf{U}^x = \mathrm{T}_x\mathcal{M}$, the manifold is called *geodesically complete*. Note that any connected compact manifold is geodesically complete. As a consequence we have that there exists a geodesic between any two points $x, y \in \mathcal{M}$ (see Lee, 2018, Lemma 6.18). For any $x, y \in \mathcal{M}$, we denote $\mathrm{Geo}_{x,y}$ the sets of geodesics γ such that $\gamma(0) = x$ and $\gamma(y) = 1$. For any $x, y \in \mathcal{M}$ we denote $\Gamma_x^y(\gamma) : \mathrm{T}_x\mathcal{M} \to \mathrm{T}_y\mathcal{M}$ the linear isomorphism such that for any $v \in \mathrm{T}_x\mathcal{M}$, $\Gamma_x^y(v) = X^v(\gamma(1))$, where $\gamma \in \mathrm{Geo}_{x,y}$. Note that for any $x \in \mathcal{M}$ there exists $\mathsf{V}^x \subset \mathcal{M}$ such that $x \in \mathsf{V}^x$ and for any $y \in \mathsf{V}^x$ we have that $|\mathrm{Geo}_{x,y}| = 1$. In this case, we denote $\Gamma_x^y = \Gamma_x^y(\gamma)$ with $\gamma \in \mathrm{Geo}_{x,y}$.

Orthogonal projection We will make repeated use of orthonormal projections on manifolds. Recall that since $\mathcal{M}$ is a closed Riemannian manifold we can use the Nash embedding theorem (Gunther, 1991). In the rest of this paragraph, we assume that $\mathcal{M}$ is a Riemannian submanifold of $\mathbb{R}^p$ for some $p \in \mathbb{N}$ such that its metric is induced by the Euclidean metric. In order to define the projection we introduce

$$\mathrm{unpp}(\mathcal{M}) = \{x \in \mathbb{R}^d : \text{ there exists a unique } \xi_x \text{ such that } \|x - \xi_x\| = d(x, \mathcal{M})\}.$$

Let $\mathcal{E}(\mathcal{M}) = \mathrm{int}(\mathrm{unpp}(\mathcal{M}))$. By Leobacher and Steinicke (2021, Theorem 1), we have that $\mathcal{M} \subset \mathcal{E}(\mathcal{M})$. We define $\tilde{p} : \mathcal{E}(\mathcal{M}) \to \mathcal{M}$ such that for any $x \in \mathcal{E}(\mathcal{M})$, $\tilde{p}(x) = \xi_x$. Using Leobacher and Steinicke (2021, Theorem 2), we have that $\tilde{p} \in \mathrm{C}^\infty(\mathbb{R}^p, \mathcal{M})$ and that for any $x \in \mathcal{M}$, $\tilde{P}(x) = \mathrm{d}\tilde{p}(x)$ is the orthogonal projection on $\mathrm{T}_x\mathcal{M}$. Since $\mathbb{R}^p$ is normal and $\mathcal{M}$ and $\mathcal{E}(\mathcal{M})^\mathrm{c}$ are closed, there exists F open such that $\mathcal{M} \subset \mathsf{F} \subset \mathcal{E}(\mathcal{M})$. Let $p \in \mathrm{C}^\infty(\mathbb{R}^p, \mathbb{R}^p)$ such that for any $x \in \mathsf{F}$, $p(x) = \tilde{p}(x)$ (given by Whitney extension theorem for instance). Finally, we define $P : \mathbb{R}^p \to \mathbb{R}^p$ such that for any $x \in \mathbb{R}^p$, $P(x) = \mathrm{d}p(x)$. Note that for any $x \in \mathcal{M}$, $P(x)$ is the orthogonal projection $\mathrm{T}_x\mathcal{M}$ and that $P \in \mathrm{C}^\infty(\mathbb{R}^p, \mathbb{R}^p)$.

S2.2 Stochastic Differential Equations on manifolds

Stratanovitch integral For reasons that will become clear in the next paragraph it is easier to define Stochastic Differential Equations (SDEs) on manifolds w.r.t the Stratanovitch integral (Kloeden and Platen, 2011, Part II, Chapter 3). We consider a filtered probability space $(\Omega, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$. Let $(\mathbf{X}_t)_{t\geq 0}$ and $(\mathbf{Y}_t)_{t\geq 0}$ be two real continuous semimartingales. We define the quadratic covariation $([\mathbf{X}, \mathbf{Y}]_t)_{t\geq 0}$ such that for any $t \geq 0$

$$[\mathbf{X}, \mathbf{Y}]_t = \mathbf{X}_t \mathbf{Y}_t - \mathbf{X}_0 \mathbf{Y}_0 - \int_0^t \mathbf{X}_s \mathrm{d}\mathbf{Y}_s - \int_0^t \mathbf{Y}_s \mathrm{d}\mathbf{X}_s.$$

We refer to Revuz and Yor (1999, Chapter IV) for more details on semimartingales and quadratic variations. We denote $[\mathbf{X}] = [\mathbf{X}, \mathbf{X}]$. In particular, we have that $([\mathbf{X}, \mathbf{Y}]_t)_{t\geq 0}$ is an adapted continuous process with finite-variation and therefore $[[\mathbf{X}, \mathbf{Y}]] = 0$. Let $(\mathbf{X}_t)_{t\geq 0}$ and $(\mathbf{Y}_t)_{t\geq 0}$ be two real continuous semimartingales, then we define the Stratanovitch integral as follows for any $t \geq 0$

$$\int_0^t \mathbf{X}_s \circ \mathrm{d}\mathbf{Y}_s = \int_0^t \mathbf{X}_s \mathrm{d}\mathbf{Y}_s + (1/2)[\mathbf{X}, \mathbf{Y}]_t.$$

In particular, denoting $(\mathbf{Z}_t^1)_{t\geq 0}$ and $(\mathbf{Z}_t^2)_{t\geq 0}$ the processes such that for any $t \geq 0$, $\mathbf{Z}_t^1 = \int_0^t \mathbf{X}_s \circ \mathrm{d}\mathbf{Y}_s$ and $\mathbf{Z}_t^2 = \int_0^t \mathbf{X}_s \mathrm{d}\mathbf{Y}_s$, we have that $[\mathbf{Z}^1] = [\mathbf{Z}^2]$. We refer to Kurtz et al. (1995) for more details on Stratanovitch integrals. Note that if for any $t \geq 0$, $\mathbf{X}_t = \int_0^t f(\mathbf{X}_s) \circ \mathrm{d}\mathbf{Y}_s$ with $\mathrm{C}^1(\mathbb{R}, \mathbb{R})$, then $[\mathbf{X}, \mathbf{Y}]_t = \int_0^t f(\mathbf{X}_s) f'(\mathbf{X}_s) \mathrm{d}\mathbf{Y}_s$. Assuming that $f \in \mathrm{C}^3(\mathbb{R}, \mathbb{R})$ we have that (Revuz and Yor, 1999, Chapter IV, Exercise 3.15)

$$f(\mathbf{X}_t) = f(\mathbf{X}_0) + \int_0^t f'(\mathbf{X}_s) \circ \mathrm{d}\mathbf{X}_s.$$

The proof relies on the fact that for any $t \geq 0$, $\mathrm{d}[\mathbf{X}, f'(\mathbf{X})]_t = f''(\mathbf{X}_t)\mathrm{d}[\mathbf{X}]_t$. This result should be compared with Itô's lemma. In particular, Stratanovitch calculus satisfies the ordinary chain rule making it a useful tool in differential geometry which makes a heavy use of diffeomorphism.

SDEs on manifolds We define semimartingales and SDEs on manifold through the lens of their actions on functions. A continuous $\mathcal{M}$-valued stochastic process $(\mathbf{X}_t)_{t\geq 0}$ is called a $\mathcal{M}$-valued semimartingale if for any $f \in \mathrm{C}^\infty(\mathcal{M})$ we have that $(f(\mathbf{X}_t))_{t\geq 0}$ is a real valued semimartingale. Let $\ell \in \mathbb{N}$, $V^{1:\ell} = \{V_i\}_{i=1}^\ell \in \mathcal{X}(\mathcal{M})^\ell$ and $Z^{1:\ell} = \{Z^i\}_{i=1}^\ell$ a collection of ℓ real-valued semimartingales. A $\mathcal{M}$-valued semimartingale $(\mathbf{X}_t)_{t\geq 0}$ is said to be the solution of $\mathrm{SDE}(V^{1:\ell}, Z^{1:\ell}, \mathbf{X}_0)$ up to a stopping τ with $\mathbf{X}_0$ a $\mathcal{M}$-valued random variable if for all $f \in \mathrm{C}^\infty(\mathcal{M})$ and $t \in [0, \tau]$ we have

$$f(\mathbf{X}_t) = f(\mathbf{X}_0) + \sum_{i=1}^\ell \int_0^t V_i(f)(\mathbf{X}_s) \circ \mathrm{d}\mathbf{Z}_s^i.$$

Since the previous SDE is defined w.r.t the Stratanovitch integral we have that if $(\mathbf{X}_t)_{t\geq 0}$ is a solution of $\mathrm{SDE}(V^{1:\ell}, Z^{1:\ell}, \mathbf{X}_0)$ and $\mathbf{\Phi} : \mathcal{M} \to \mathcal{N}$ is a diffeomorphism then $(\mathbf{\Phi}(\mathbf{X}_t))_{t\geq 0}$ is a solution of $\mathrm{SDE}(\mathbf{\Phi}_\star V^{1:\ell}, Z^{1:\ell}, \mathbf{\Phi}(\mathbf{X}_0))$, where $\mathbf{\Phi}_\star$ is the pushforward operation (see Hsu, 2002, Proposition 1.2.4). Because the vector fields $\{V_i\}_{i=1}^\ell$ are smooth we have that for any $\ell \in \mathbb{N}$, $V^{1:\ell} = \{V_i\}_{i=1}^\ell \in \mathcal{X}(\mathcal{M})^\ell$ and $Z^{1:\ell} = \{Z^i\}_{i=1}^\ell$ a collection of ℓ real-valued semimartingales, there exists a unique solution to $\mathrm{SDE}(V^{1:\ell}, Z^{1:\ell}, \mathbf{X}_0)$ (see Hsu, 2002, Theorem 1.2.9).

S2.3 Frame bundle and orthonormal frame bundle

We now introduce the concepts of frame bundle and orthonormal bundle over the manifold $\mathcal{M}$. These concepts are useful to define stochastic processes on $\mathcal{M}$ using Euclidean stochastic processes. In particular, we will see that a Brownian motion on the manifold can be linked to the Euclidean Brownian motion using the orthonormal bundle. For any $x \in \mathcal{M}$, a frame at x is an isomorphism $f : \mathbb{R}^d \to \mathrm{T}_x\mathcal{M}$. Note that f is equivalent to the choice of a basis in $\mathrm{T}_x\mathcal{M}$. We denote $\mathrm{F}_x\mathcal{M}$ the set of frames at p. The frame bundle denoted $\mathrm{F}\mathcal{M}$ is given by $\mathrm{F}\mathcal{M} = \sqcup_{x\in\mathcal{M}} \mathrm{F}_x\mathcal{M}$. The frame bundle can be given a smooth structure and is therefore a $d + d^2$-dimensional manifold. Similarly, for any $x \in \mathcal{M}$, an orthonormal frame at x is a linear isometry $f : \mathbb{R}^d \to \mathrm{T}_x\mathcal{M}$. Note that f is equivalent to the choice of an orthonormal basis in $\mathrm{T}_x\mathcal{M}$. We denote $\mathrm{O}_p\mathcal{M}$ the set of orthonormal frames at p. The orthonormal frame bundle denoted $\mathrm{O}\mathcal{M}$ is given by $\mathrm{O}\mathcal{M} = \sqcup_{x\in\mathcal{M}} \mathrm{O}_x\mathcal{M}$. The orthonormal frame bundle can be given a smooth structure and is therefore a $d + d(d-1)/2$-dimensional manifold. We denote $\pi : \mathrm{F}\mathcal{M} \to \mathcal{M}$ the smooth projection such that for any $u = (x, f) \in \mathrm{F}\mathcal{M}$, $\pi(u) = x$. Note that the restriction of π to the orthonormal bundle is also smooth. Frame bundles and orthonormal bundles are primary examples of principal bundles and we refer to Kolár et al. (2013) for more details.

One key element of frame bundles and orthonormal bundles is their link with the connections on $\mathcal{M}$. Let $u = (x, f) \in \mathrm{F}\mathcal{M}$ and $U \in \mathrm{T}_u FM$. U is said to be vertical if there exists a smooth curve $u : [0,1] \to \mathrm{F}\mathcal{M}$ such that for any $t \in [0,1]$, $\pi(u(t)) = x$ and $\dot{u}(0) = U$. We say that U is tangent to the fibre $\mathrm{F}_{\pi(u)}\mathcal{M}$. The space of vertical tangent vectors is called the vertical space and is denoted $\mathrm{V}_u\mathrm{F}\mathcal{M}$. We have that $\dim(\mathrm{V}_u\mathrm{F}\mathcal{M}) = d^2$. We now define the horizontal space as follows. Let $u : [0,1] \to \mathrm{F}\mathcal{M}$ be a smooth curve. We say that $u = (f, x)$ is horizontal if for any $t \in [0,1]$ and $i \in \{1, \dots, d\}$, $\nabla_{\dot{x}}(fe_i)(t) = 0$, where $\{e_i\}_{i=1}^d$ is the canonical basis of $\mathbb{R}^d$. In other words, the horizontal curve corresponds to the parallel transport of a frame along a smooth curve in $\mathcal{M}$. Let $u = (x, f) \in \mathrm{F}\mathcal{M}$ and $U \in \mathrm{T}_u\mathrm{F}M$. U is said to be horizontal if there exists a smooth horizontal curve $u : [0,1] \to \mathrm{F}\mathcal{M}$ such that $\dot{u}(0) = U$. The space of horizontal tangent vectors is called the horizontal space and is denoted $\mathrm{H}_u\mathrm{F}\mathcal{M}$. Let $v \in \mathbb{R}^d$, we define the vector field $H_v \in \mathcal{X}(\mathrm{F}\mathcal{M})$ such that for any $u \in \mathrm{F}\mathcal{M}$, $H_v(u) = \dot{u}(0)$ with $\gamma = (x, f) : [0,1] \to \mathrm{F}\mathcal{M}$ a smooth curve on $\mathrm{F}\mathcal{M}$ such that $\dot{x}(0) = e(0)v$ and $\gamma(0) = u$. The existence of H_v for any $v \in \mathbb{R}^d$ is discussed in Kobayashi and Nomizu (1963, p.69-70) and H_v is called the horizontal lift of v. For any $i \in \{1, \dots, d\}$ we denote $H_i = H_{e_i}$ where $\{e_i\}_{i=1}^d$ is the canonical basis of $\mathbb{R}^d$. In particular, since any horizontal curve is entirely specified by $\gamma(0) = (x(0), f(0))$ and $\dot{x}(0)$, we get that $\dim(\mathrm{H}_u\mathrm{F}\mathcal{M}) = d$ for any $u \in \mathrm{F}\mathcal{M}$.

Consider a connection ∇ on $\mathcal{M}$. Note that for any $u = (x, f) \in \mathrm{F}\mathcal{M}$, we have $\mathrm{T}_u\mathrm{F}\mathcal{M} = \mathrm{T}_u\mathcal{M} \oplus \mathrm{V}_u\mathrm{F}\mathcal{M}$. In local coordinates $\{x_i\}_{i=1}^d$, we denote $\{X_i\}_{i=1}^d$ a basis of $\mathrm{T}_x\mathcal{M}$. For any $j \in \{1, \dots, j\}$, there exist $\{f_{i,j}\}_{i=1}^d$ such that $fe_j = \sum_{i=1}^d f_{i,j} X_i$ (note that $\{f_{i,j}\}_{1 \le i,j \le d}$ can be interpreted as the matrix transforming a vector of $\mathbb{R}^d$ into a vector of $\mathrm{T}_x\mathcal{M}$ expressed in the basis $\{X_i\}_{i=1}^d$). In particular, we have that $\{x_k, f_{i,j}\}_{1 \le i,j,k \le d}$ are local coordinates for $\mathrm{F}\mathcal{M}$. We denote by $\{X_k, X_{i,j}\}_{1 \le i,j,k \le d}$ the associated basis in $\mathrm{T}_u\mathrm{F}\mathcal{M}$ for any $u \in \mathsf{U}$, where U is an open subset of $\mathrm{F}\mathcal{M}$ on which the local coordinates are well-defined. Leveraging properties of parallel transport, we have that for any $j \in \{1, \dots, d\}$ and $u \in \mathsf{U}$

$$H_j(u) = \textstyle\sum_{i=1}^d f_{i,j} X_i - \sum_{\ell,m=1}^d \{\sum_{i,k=1}^d f_{i,j} f_{k,m} \Gamma_{i,k}^\ell\} X_{\ell,m}, \tag{S2}$$

where we recall that $\{\Gamma_{i,j}^k\}_{1 \le i,j,k \le d}$ are the Christoffel symbols of the connection in local coordinates. In particular, it is clear that for any $u \in \mathrm{F}\mathcal{M}$, $\{H_i(u)\}_{i=1}^d$ is a basis of $\mathrm{H}_u\mathrm{F}\mathcal{M}$ and that $\mathrm{H}_u\mathrm{F}\mathcal{M} \cap \mathrm{V}_u\mathrm{F}\mathcal{M} = \{0\}$, hence $\mathrm{T}_u\mathrm{F}\mathcal{M} = \mathrm{H}_u\mathrm{F}\mathcal{M} \oplus \mathrm{V}_u\mathrm{F}\mathcal{M}$. Using Equation (S2) we have that the horizontal space is entirely defined by the connection ∇. Reciprocally, any smooth linear complement of the vertical space gives rise to a connection (see Kolář et al., 2013, Section 11.11).

We now illustrate how we can go from a smooth curve on $\mathcal{M}$ (equipped with a connection ∇) to a smooth curve on $\mathbb{R}^d$. First, let $x : [0,1] \to \mathcal{M}$ be a smooth curve on manifold. Define $f(0) \in \mathrm{F}_{x(0)}\mathcal{M}$ and consider $u : [0,1] \to \mathrm{F}\mathcal{M}$ the smooth horizontal curve associated with x and starting frame $f(0)$. Now consider the antidevelopment of u given by the smooth curve $z : [0,1] \to \mathbb{R}^d$ such that for any $t \in [0,t]$

$$z(t) = \textstyle\int_0^t f(s)^{-1} \dot{x}(s) \mathrm{d}s.$$

We now show how a smooth curve on $\mathbb{R}^d$ gives rise to a smooth curve in $\mathcal{M}$. First, note that for any $t \in [0,1]$, we have that $\dot{u}(t) = \sum_{i=1}^d H_i(u(t)) \dot{z}_i(t)$. Hence, specifying $u(0)$ any smooth curve z on $\mathbb{R}^d$ is associated to a smooth curve on $\mathrm{F}\mathcal{M}$. We obtain a smooth curve on $\mathcal{M}$ by considering $x = \pi(u)$. In the next section, we present similar ideas when smooth curves are replaced by semimartingales.

S2.4 Horizontal lift and stochastic development

We are now ready to present the notion of horizontal semimartingale, which is key to draw the link between semimartingales on $\mathcal{M}$ and semimartingales on $\mathbb{R}^d$. We follow the presentation of Hsu (2002, Section 2.3). Again, we consider a filtered probability space $(\Omega, (\mathcal{F}_t)_{t \ge 0}, \mathbb{P})$. All the semimartingales we consider are defined w.r.t this filtered probability space. We assume that the manifold $\mathcal{M}$ is equipped with a connection ∇.

Definition S8 (Stochastic development)**.** *Let $(\mathbf{Z}_t^{1:d})_{t \ge 0} = \{(\mathbf{Z}_t^i)_{t \ge 0}\}_{i=1}^d$ be a collection of real-valued semimartingales. Let $(\mathbf{U}_t)_{t \ge 0}$ be the* $\mathrm{F}\mathcal{M}$ *semimartingale solution of* $\mathrm{SDE}(H^{1:d}, \mathbf{Z}^{1:d}, \mathbf{U}_0)$ *with $H^{1:d} = \{H_i\}_{i=1}^d$. $(\mathbf{U}_t)_{t \ge 0}$ is called the* stochastic development *of $\mathbf{Z}^{1:d}$ on* $\mathrm{F}\mathcal{M}$. *Similarly, the $\mathcal{M}$-valued semimartingale $(\mathbf{X}_t)_{t \ge 0} = (\pi(\mathbf{U}_t))_{t \ge 0}$ is called the* stochastic development *of $\mathbf{Z}^{1:d}$ on $\mathcal{M}$.*

The previous definition allows to transfer a semimartingale on $\mathbb{R}^d$ to a semimartingale on $\mathcal{M}$ in an *intrinsic* manner. Reciprocally, we also aim at transferring a semimartingale on $\mathcal{M}$ to a semimartingale on $\mathbb{R}^d$.

Definition S9 (Horizontal lift and antivelopment)**.** *Let $(\mathbf{X}_t)_{t \ge 0}$ be a $\mathcal{M}$-valued semimartingale. If there exist a $\mathrm{F}\mathcal{M}$-valued semimartingale $(\mathbf{U}_t)_{t \ge 0}$ and $(\mathbf{Z}_t^{1:d})_{t \ge 0} = \{(\mathbf{Z}_t^i)_{t \ge 0}\}_{i=1}^d$ a collection of real-valued semimartingales such that $(\mathbf{X}_t)_{t \ge 0} = (\pi(\mathbf{U}_t))_{t \ge 0}$ and $(\mathbf{U}_t)_{t \ge 0}$ is solution of* $\mathrm{SDE}(H^{1:d}, \mathbf{Z}^{1:d}, \mathbf{U}_0)$ *with $H^{1:d} = \{H_i\}_{i=1}^d$ then $(\mathbf{U}_t)_{t \ge 0}$ is called the* horizontal lift *of $(\mathbf{X}_t)_{t \ge 0}$ and $(\mathbf{Z}_t^{1:d})_{t \ge 0}$ the* antidevelopment *of $(\mathbf{X}_t)_{t \ge 0}$.*

The existence of an horizontal lift and an antidevelopment is not trivial. Considering the Nash embedding theorem (see for example Gunther, 1991), it is possible to show the existence and uniqueness of these processes (up to initialization). Without loss of generality, we can then assume that $\mathcal{M} \subset \mathbb{R}^p$ and for any

$x \in \mathcal{M}$, $\mathrm{T}_x\mathcal{M} \subset \mathbb{R}^p$ with $p \geq d(d+1)/2$ (and $p \leq \max(d(d+5)/2, d(d+3)/2+5)$). For any $x \in \mathcal{M}$, we denote $P(x): \mathbb{R}^p \to \mathrm{T}_x\mathcal{M}$ the projection operator. In addition for any $x \in \mathcal{M}$, we denote $\{P_i(x)\}_{i=1}^p = \{P(x)e_i\}_{i=1}^p$, where $\{e_i\}_{i=1}^p$ is the canonical basis of $\mathbb{R}^p$. Note that $\{P_i\}_{i=1}^p \in \mathcal{X}(\mathcal{M})^p$. In addition for any $x \in \mathcal{M}$ we denote $\{x^i\}_{i=1}^p$ its coordinates in $\mathbb{R}^p$, *i.e.* for any $i \in \{1, \dots, p\}$, $x^i = \langle x, e_i \rangle$. In particular, if $(\mathbf{X}_t)_{t\geq 0}$ is a $\mathcal{M}$-valued process then for any $i \in \{1, \dots, p\}$, $(\mathbf{X}_t^i)_{t\geq 0} = (\langle \mathbf{X}_t, e_i \rangle)_{t \geq 0}$ is a real-valued process. If $(\mathbf{X}_t)_{t\geq 0}$ is a $\mathcal{M}$-valued semimartingale then it is the solution of $\mathrm{SDE}(\{P_i\}_{i=1}^p, \{\mathbf{X}^i\}_{i=1}^p, \mathbf{X}_0)$ (see Hsu, 2002, Lemma 2.3.3). Then, a candidate for the horizontal lift of $(\mathbf{X}_t)_{t\geq 0}$ is given by $(\mathbf{U}_t)_{t\geq 0} = (\mathbf{X}_t, \mathbf{E}_t)_{t\geq 0}$ solution of $\mathrm{SDE}(\{P_i^\star\}_{i=1}^p, \{\mathbf{X}^i\}_{i=1}^p, \mathbf{U}_0)$, where for any $i \in \{1, \dots, p\}$, $P_i^\star(u) = H_{f^{-1}P_i(\pi(u))}(u)$ and $\mathbf{X}_0 = \pi(\mathbf{U}_0)$. We have that $(\mathbf{U}_t)_{t\geq 0}$ is the stochastic development of $\{(\mathbf{Z}_t^i)_{t\geq 0}\}_{i=1}^d$ where for any $t \geq 0$, $\mathbf{Z}_t = \sum_{i=1}^p \int_0^t \mathbf{E}_s^{-1} P_i(\mathbf{X}_s) \circ \mathrm{d}\mathbf{X}_s^i$ (see Hsu, 2002, Theorem 2.3.4). Finally, we have that given $\mathbf{U}_0$, $(\mathbf{U}_t)_{t\geq 0}$ is the unique horizontal lift of $(\mathbf{X}_t)_{t\geq 0}$ and $(\mathbf{Z}_t)_{t\geq 0}$ is the unique antidevelopment of $(\mathbf{X}_t)_{t\geq 0}$ (see Hsu, 2002, Theorem 2.3.5).

S2.5 Brownian motion on manifolds

In this section, we introduce the notion of Brownian motion on manifolds. We derive some of its basic convergence properties and provide alternative definitions (stochastic development, isometric embedding, random walk limit). These alternative definitions are the basis for our alternative methodologies to sample from the time-reversal. To simplify our discussion, we assume that $\mathcal{M}$ is a connected compact Riemannian manifold equipped with the Levi-Civita connection ∇. We denote $p_{\mathrm{ref}}b$ the Haussdorff measure of the manifold (which coincides with the measure associated with the Riemannian volume form (see Federer, 2014, Theorem 2.10.10) and $p_{\mathrm{ref}} = p_{\mathrm{ref}}b/p_{\mathrm{ref}}(\mathcal{M})$ the associated probability measure.

Gradient, divergence and Laplace operators Let $f \in \mathrm{C}^\infty(\mathcal{M})$. We define $\nabla f \in \mathcal{X}(\mathcal{M})$ such that for any $X \in \mathcal{X}(\mathcal{M})$ we have $\langle X, \nabla f \rangle_{\mathcal{M}} = X(f)$. Let $\{X_i\}_{i=1}^d \in \mathcal{X}(\mathcal{M})^d$ such that for any $x \in \mathcal{M}$, $\{X_i(x)\}_{i=1}^d$ is an orthonormal basis of $\mathrm{T}_x\mathcal{M}$. Then, we define $\mathrm{div}: \mathcal{X}(\mathcal{M}) \to \mathrm{C}^\infty(\mathcal{M})$ (linear) such that for any $X \in \mathcal{X}(\mathcal{M})$, $\mathrm{div}(X) = \sum_{i=1}^d \langle \nabla_{X_i} X, X_i \rangle_{\mathcal{M}}$. The following Stokes formula (also called divergence theorem, see Lee (2018, p.51)) holds for any $f \in \mathrm{C}^\infty(\mathcal{M})$ and $X \in \mathcal{X}(\mathcal{M})$, $\int_M \mathrm{div}(X)(x) f(x) \mathrm{d}p_{\mathrm{ref}}(x) = -\int_M X(f)(x) \mathrm{d}p_{\mathrm{ref}}(x)$. Let $X = \sum_{i=1}^d a_i X_i$ in local coordinates. Using the Stokes formula and the definition of the gradient we get that in local coordinates

$$\nabla f = \textstyle\sum_{i,j=1}^d g^{i,j} \partial_i f X_j, \qquad \mathrm{div}(X) = \det(G)^{-1/2} \textstyle\sum_{i=1}^d \partial_i (\det(G)^{1/2} a_i).$$

The Laplace-Beltrami operator is given by $\Delta_{\mathcal{M}}: \mathrm{C}^\infty(M) \to \mathrm{C}^\infty(M)$ and for any $f \in \mathrm{C}^\infty(M)$ by $\Delta_{\mathcal{M}}(f) = \mathrm{div}(\mathrm{grad}(f))$. In local coordinates we obtain $\Delta_{\mathcal{M}}(f) = \det(G)^{-1/2} \sum_{i=1}^d \partial_i (\det(G)^{1/2} \sum_{j=1}^d g^{i,j} \partial_j f)$. Using the Nash isometric embedding theorem (Gunther, 1991) we will see that $\Delta_{\mathcal{M}}$ can always be written as a sum of squared operators. However, this result requires an *extrinsic* point of view as it relies on the existence of projection operators. In contrast, if we consider the orthonormal bundle $\mathrm{O}\mathcal{M}$ we can define the Laplace-Bochner operator $\Delta_{\mathrm{O}\mathcal{M}}: \mathrm{C}^\infty(\mathrm{O}\mathcal{M}) \to \mathrm{C}^\infty(\mathrm{O}\mathcal{M})$ as $\Delta_{\mathrm{O}\mathcal{M}} = \sum_{i=1}^d H_i^2$, where we recall that for any $i \in \{1, \dots, d\}$, H_i is the horizontal lift of e_i. In this case, $\Delta_{\mathrm{O}\mathcal{M}}$ is a sum of squared operators and we have that for any $f \in \mathrm{C}^\infty(\mathcal{M})$, $\Delta_{\mathrm{O}\mathcal{M}}(f \circ \pi) = \Delta_{\mathcal{M}}(f)$ (see Hsu, 2002, Proposition 3.1.2). Being able to express the various Laplace operators as a sum of squared operators is key to express the associated diffusion process as the solution of an SDE.

Alternatives definitions of Brownian motion We are now ready to define a Brownian motion on the manifold $\mathcal{M}$. Using the Laplace-Beltrami operator, we can introduce the Brownian motion through the lens of diffusion processes.

Definition S10 (Brownian motion)**.** *Let $(\mathbf{B}_t^{\mathcal{M}})_{t\geq 0}$ be a $\mathcal{M}$-valued semimartingale. $(\mathbf{B}_t^{\mathcal{M}})_{t\geq 0}$ is a Brownian motion on $\mathcal{M}$ if for any $f \in \mathrm{C}^\infty(\mathcal{M})$, $(\mathbf{M}_t^f)_{t\geq 0}$ is a local martingale where for any $t \geq 0$*

$$\mathbf{M}_t^f = f(\mathbf{B}_t^{\mathcal{M}}) - f(\mathbf{B}_0^{\mathcal{M}}) - (1/2) \textstyle\int_0^t \Delta_{\mathcal{M}} f(\mathbf{B}_s^{\mathcal{M}}) \mathrm{d}s.$$

Note that this definition is in accordance with the definition of the Brownian motion as a diffusion process in the Euclidean space $\mathbb{R}^d$, since in this case $\Delta_{\mathcal{M}} = \Delta$. As emphasized in the previous section any semimartingale on $\mathcal{M}$ can be associated to a process on $\mathrm{F}\mathcal{M}$ (or $\mathrm{O}\mathcal{M}$) and a process on $\mathbb{R}^d$. The proof of the following result can be found in Hsu (2002, Propositions 3.2.1 and 3.2.2).

Proposition S11 (Intrinsic view of Brownian motion)**.** *Let $(\mathbf{B}_t^{\mathcal{M}})_{t\geq 0}$ be a $\mathcal{M}$-valued semimartingales. Then $(\mathbf{B}_t^{\mathcal{M}})_{t\geq 0}$ is a Brownian motion on $\mathcal{M}$ if and only on the following conditions hold:*

a) *The horizontal lift* $(\mathbf{U}_t)_{t\geq 0}$ *is a* $\Delta_{\mathrm{O}\mathcal{M}}/2$ *diffusion process, i.e. for any* $f \in \mathrm{C}^\infty(\mathrm{O}\mathcal{M})$*, we have that* $(\mathbf{M}_t^f)_{t\geq 0}$ *is a local martingale where for any* $t \geq 0$

$$\mathbf{M}_t^f = f(\mathbf{U}_t) - f(\mathbf{U}_0) - (1/2)\int_0^t \Delta_{\mathrm{O}\mathcal{M}} f(\mathbf{U}_s)\mathrm{d}s.$$

b) *The stochastic antidevelopment of* $(\mathbf{B}_t^{\mathcal{M}})_{t\geq 0}$ *is a* $\mathbb{R}^d$*-valued Brownian motion* $(\mathbf{B}_t)_{t\geq 0}$*.*

In particular the previous proposition provides us with an *intrisic* way to sample the Brownian motion on $\mathcal{M}$ with initial condition $\mathbf{B}_0^{\mathcal{M}}$. First sample $(\mathbf{U}_t)_{t\geq 0}$ solution of $\mathrm{SDE}(H^{1:d}, \mathbf{B}^{1:d}, \mathbf{U}_0)$ with $H^{1:d} = \{H_i\}_{i=1}^d$ and $\pi(\mathbf{U}_0) = \mathbf{B}_0^{\mathcal{M}}$ and $\mathbf{B}^{1:d}$ the Euclidean d-dimensional Brownian motion. Then, we recover the $\mathcal{M}$-valued Brownian motion $(\mathbf{B}_t^{\mathcal{M}})_{t\geq 0}$ upon letting $(\mathbf{B}_t^{\mathcal{M}})_{t\geq 0} = (\pi(\mathbf{U}_t))_{t\geq 0}$.

We now consider an *extrinsic* approach to the sampling of Brownian motions on $\mathcal{M}$. Using the Nash embedding theorem (Gunther, 1991), there exists $p \in \mathbb{N}$ such that without loss of generality we can assume that $\mathcal{M} \subset \mathbb{R}^p$. For any $x \in \mathcal{M}$, we denote $P(x) : \mathbb{R}^p \to \mathrm{T}_x\mathcal{M}$ the projection operator. In addition for any $x \in \mathcal{M}$, we denote $\{P_i(x)\}_{i=1}^p = \{P(x)e_i\}_{i=1}^p$, where $\{e_i\}_{i=1}^p$ is the canonical basis of $\mathbb{R}^p$. For any $i \in \{1, \dots, p\}$, we smoothly extend P_i to $\mathbb{R}^p$. In this case, we have the following proposition (Hsu, 2002, Theorem 3.1.4):

Proposition S12 (Extrinsic view of Brownian motion)**.** *For any* $f \in \mathrm{C}^\infty(\mathcal{M})$ *we have that* $\Delta_M(f) = \sum_{i=1}^p P_i(P_i(f))$*. Hence, we have that* $(\mathbf{B}_t^{\mathcal{M}})_{t\geq 0}$ *solution of* $\mathrm{SDE}(\{P_i\}_{i=1}^p, \mathbf{B}^{1:p}, \mathbf{B}_0^{\mathcal{M}})$ *with* $\mathbf{B}_0^{\mathcal{M}}$ *a* $\mathcal{M}$*-valued random variable and* $\mathbf{B}^{1:p}$ *a* $\mathbb{R}^p$*-valued Brownian motion.*

The second part of this proposition, stems from the fact that any solution of $\mathrm{SDE}(\{V_i\}_{i=1}^\ell, \mathbf{B}^{1:\ell}, \mathbf{X}_0)$, where $\mathbf{X}_0$ is a $\mathcal{M}$-valued random variable and $\mathbf{B}^{1:\ell}$ a $\mathbb{R}^\ell$-valued Brownian motion is a diffusion process with generator $\mathcal{A}$ such that for any $f \in \mathrm{C}^\infty(\mathcal{M})$, $\mathcal{A}(f) = \sum_{i=1}^\ell V_i(V_i(f))$. The *extrinsic* approach is particularly convenient since the SDE appearing in Proposition S12 can be seen as an SDE on the Euclidean space $\mathbb{R}^p$.

We finish this paragraph, by investigating the behavior of the Brownian motion in local coordinates. For simplicity, we assume here that we have access to a system of global coordinates. In the case where the coordinates are strictly local then we refer to Ikeda and Watanabe (1989, Chapter 5, Theorem 1) for a construction of a global solution by patching local solutions. We denote $\{X_k, X_{i,j}\}_{1\leq i,j,k\leq d}$ such that for any $u \in \mathrm{F}\mathcal{M}$, $\{X_k(u), X_{i,j}(u)\}_{1\leq i,j,k\leq d}$ is a basis of $\mathrm{T}_u\mathrm{F}\mathcal{M}$, similarly as in the previous section. Using Equation (S2) we get that $(\mathbf{U}_t)_{t\geq 0} = (\{\mathbf{X}_t^k, \mathbf{E}_t^{i,j}\}_{1\leq i,j,k\leq d})$ obtained in Proposition S11 is given in the global coordinates for any $i, j, k \in \{1, \dots, d\}$ by

$$\mathrm{d}\mathbf{X}_t^k = \sum_{j=1}^d \mathbf{E}_t^{k,j} \circ \mathrm{d}\mathbf{B}_t^k, \qquad \mathrm{d}\mathbf{E}_t^{i,j} = -\sum_{n=1}^d \{\sum_{\ell,m=1}^d \mathbf{E}_t^{\ell,n}\mathbf{E}_t^{m,j}\Gamma_{\ell,m}^i(\mathbf{X}_t)\} \circ \mathrm{d}\mathbf{B}_t^n.$$

By definition of the Stratanovitch integral we have that for any $k \in \{1, \dots, d\}$

$$\mathrm{d}\mathbf{X}_t^k = \sum_{j=1}^d \{\mathbf{E}_t^{k,j}\mathrm{d}\mathbf{B}_t^k + (1/2)\mathrm{d}[\mathbf{E}_t^{k,j}, \mathbf{B}_t^j]_t\}.$$

Let $(\mathbf{M}_t)_{t\geq 0} = (\{\mathbf{M}_t^k\}_{k=1}^d)_{t\geq 0}$ such that for any $t \geq 0$ and $k \in \{1, \dots, d\}$ $\mathbf{M}_t^k = \sum_{j=1}^d \int_0^t \mathbf{E}_t^{k,j}\mathrm{d}\mathbf{B}_t^k$. We obtain that $\mathrm{d}\mathbf{M}_t = G(\mathbf{X}_t)^{-1/2}\mathrm{d}\mathbf{B}_t$ for some d-dimensional Brownian motion $(\mathbf{B}_t)_{t\geq 0}$, using Lévy's characterization of Brownian motion. In addition, we have that for any $k, j \in \{1, \dots, d\}$

$$[\mathbf{E}^{k,j}, \mathbf{B}^j]_t = -\sum_{\ell,m=1}^d \int_0^t \mathbf{E}_t^{\ell,j}\mathbf{E}_t^{m,j}\Gamma_{\ell,m}^k(\mathbf{X}_t)\mathrm{d}t$$

Hence, using this result and the fact that $\sum_{j=1}^d \mathbf{E}_t^{\ell,j}\mathbf{E}_t^{m,j} = g^{\ell,m}(\mathbf{X}_t)$, we get that for any $k \in \{1, \dots, d\}$

$$\mathrm{d}\mathbf{X}_t^k = -(1/2)\sum_{\ell,m=1}^d g^{\ell,m}(\mathbf{X}_t)\Gamma_{\ell,m}^k(\mathbf{X}_t)\mathrm{d}t + (G(\mathbf{X}_t)^{-1/2}\mathrm{d}\mathbf{B}_t)^k.$$

Note that this result could also have been obtained using the expression of the Laplace-Beltrami in local coordinates.

Brownian motion and random walks In the previous paragraph we consider three SDEs to obtain a Brownian motion on $\mathcal{M}$ (stochastic development, isometric embedding and local coordinates). In this section, we summarize results from Jørgensen (1975) establishing the limiting behavior of Geodesic Random Walks (GRWs) when the stepsize of the random walk goes to 0. This will be of particular interest when considering the time-reversal process. We start by defining the geodesic random walk on $\mathcal{M}$, following Jørgensen (1975, Section 2).

Let $\{\nu_x\}_{x\in\mathcal{M}}$ such that for any $x \in \mathcal{M}$, $\nu_x : \mathcal{B}(\mathrm{T}_x\mathcal{M}) \to [0,1]$ with $\nu_x(\mathrm{T}_x\mathcal{M}) = 1$, *i.e.* for any $x \in \mathcal{M}$, ν_x is a probability measure on $\mathrm{T}_x\mathcal{M}$. Assume that for any $x \in \mathcal{M}$, $\int_{\mathcal{M}} \|v\|^3 \mathrm{d}\nu_x(v) < +\infty$. In addition assume that there exists $\mu^{(1)} \in \mathcal{X}(\mathcal{M})$ and $\mu^{(2)} \in \mathcal{X}^2(\mathcal{M})$, where $\mathcal{X}^2(\mathcal{M})$ is the section $\Gamma(\mathcal{M}, \sqcup_{x\in\mathcal{M}}\mathcal{L}(\mathrm{T}_x\mathcal{M}))$, such that for any $x \in \mathcal{M}$, $\int_{\mathcal{M}} v\mathrm{d}\nu_x(v) = \mu^{(1)}(x)$ and $\int_{\mathcal{M}} v \otimes v\mathrm{d}\nu_x(v) = \mu^{(2)}(x)$. In addition, we assume that for any $x \in \mathcal{M}$, $\Sigma(x) = \mu^{(2)}(x) - \mu^{(1)}(x) \otimes \mu^{(1)}(x)$ is strictly positive definite and that there exists $\mathtt{L} \geq$ such that for any $x, y \in \mathcal{M}$, $\|\nu_x - \nu_y\|_{\mathrm{TV}} \leq \mathtt{L}d(x,y)$. Where we have that for any $\nu_1 \in \mathcal{P}(\mathrm{T}_x\mathcal{M})$ and $\nu_2 \in \mathcal{P}(\mathrm{T}_y\mathcal{M})$,

$$\|\nu_x - \nu_y\|_{\mathrm{TV}} = \sup\{\nu_1[f] - \Gamma_x^y(\gamma)_\#\nu_2[f] \ : \ \gamma \in \mathrm{Geo}_{x,y}, \ f \in \mathrm{C}(\mathrm{T}_x\mathcal{M})\}.$$

Note that if $d(x,y) \leq \varepsilon$ then for some $\varepsilon > 0$ we have that $|\mathrm{Geo}_{x,y}| = 1$.

Definition S13 (Geodesic random walk)**.** *Let X_0 be a $\mathcal{M}$-valued random variable. For any $\gamma > 0$, we define $(\mathbf{X}_t^\gamma)_{t\geq 0}$ such that $\mathbf{X}_0^\gamma = X_0$ and for any $n \in \mathbb{N}$ and $t \in [0,\gamma]$, $\mathbf{X}_{n\gamma+t} = \exp_{\mathbf{X}_{n\gamma}}[t\gamma\{\mu_n + (1/\sqrt{\gamma})(V_n - \mu_n)\}]$, where $(V_n)_{n\in\mathbb{N}}$ is a sequence of random variables in such that for any $n \in \mathbb{N}$, V_n has distribution $\nu_{\mathbf{X}_{n\gamma}}$ conditionally to $\mathbf{X}_{n\gamma}$.*

For any $\gamma > 0$, the process $(X_n^\gamma)_{n\in\mathbb{N}} = (\mathbf{X}_{n\gamma}^\gamma)_{n\in\mathbb{N}}$ is called a geodesic random walk. In particular, for any $\gamma > 0$ we denote $(\mathrm{R}_n^\gamma)_{n\in\mathbb{N}}$ the sequence of Markov kernels such that for any $n \in \mathbb{N}$, $x \in \mathcal{M}$ and $\mathsf{A} \in \mathcal{B}(\mathcal{M})$ we have that $\delta_x\mathrm{R}(\mathsf{A}) = \mathbb{P}(X_n^\gamma \in \mathsf{A})$, with $X_0^\gamma = x$. The following theorem establishes that the limiting dynamics of a geodesic random walk is associated with a diffusion process on $\mathcal{M}$ whose coefficients only depends on the properties of ν (see Jørgensen, 1975, Theorem 2.1).

Theorem S14 (Convergence of geodesic random walks)**.** *For any $t \geq 0$, $f \in \mathrm{C}(\mathcal{M})$ and $x \in \mathcal{M}$ we have that $\lim_{\gamma\to 0} \|\mathrm{R}_\gamma^{\lceil t/\gamma\rceil}[f] - \mathrm{P}_t[f]\|_\infty = 0$, where $(\mathrm{P}_t)_{t\geq 0}$ is the semi-group associated with the infinitesimal generator $\mathcal{A} :\ \mathrm{C}^\infty(\mathcal{M}) \to \mathrm{C}^\infty(\mathcal{M})$ given for any $f \in \mathrm{C}^\infty(\mathcal{M})$ by $\mathcal{A}(f) = \langle p_{\mathrm{ref}}^1, \nabla f\rangle_{\mathcal{M}} + (1/2)\langle \Sigma, \nabla^2 f\rangle_{\mathcal{M}}$.*

In particular if $\mu^{(1)} = 0$ and $\mu^{(2)} = \mathrm{Id}$ then the random walk converges towards a Brownian motion on $\mathcal{M}$ in the sense of the convergence of semi-groups. For any $x \in \mathcal{M}$ in local coordinates we have that $\Phi_\#\nu_x$ has zero mean and covariance matrix $G(x)$, where Φ is a local chart around x and $G(x) = (g_{i,j}(x))_{1\leq i,j\leq d}$ the coordinates of the metric in that chart.

Convergence of Brownian motion We finish this section with a few considerations regarding the convergence of the Brownian motion on $\mathcal{M}$. Since we have assumed that $\mathcal{M}$ is compact we have that there exist $(\Phi_k)_{k\in\mathbb{N}}$ an orthonormal basis of $\Delta_{\mathcal{M}}$ in $\mathrm{L}^2(p_{\mathrm{ref}})$, $(\lambda_k)_{k\in\mathbb{N}}$ such that for any $i,j \in \mathbb{N}$, $i \leq j$, $\lambda_i \leq \lambda_j$ and $\lambda_0 = 0$, $\Phi_0 = 1$ and for any $k \in \mathbb{N}$, $\Delta_{\mathcal{M}}\Phi_k = -\lambda_k\Phi_k$. For any $t \geq 0$ and $x, y \in \mathcal{M}$, $p_t(x,y) = \sum_{k\in\mathbb{N}} \exp[-\lambda_k t]\Phi_k(x)\Phi_k(y)$ where for any $f \in \mathrm{C}^\infty$ we have

$$\mathbb{E}[f(\mathbf{B}_t^{\mathcal{M},x})] = \textstyle\int_{\mathcal{M}} p_t(x,y)f(y)\mathrm{d}p_{\mathrm{ref}}(y),$$

where $(\mathbf{B}_t^{\mathcal{M},x})_{t\geq 0}$ is the Brownian motion on $\mathcal{M}$ with $\mathbf{B}_0^{\mathcal{M},x} = x$ and p_{ref} is the probability measure associated with the Haussdorff measure on $\mathcal{M}$. we also have the following result (see Urakawa, 2006, Proposition 2.6).

Proposition S15 (Concergence of Brownian motion)**.** *For any $t > 0$, P_t admits a density p_t w.r.t p_{ref} and $p_{\mathrm{ref}}\mathrm{P}_t = p_{\mathrm{ref}}$, i.e. p_{ref} is an invariant measure for $(\mathrm{P}_t)_{t\geq 0}$. In addition, if there exists $C, \alpha \geq 0$ such that for any $t \in (0,1]$, $p_t(x,x) \leq Ct^{-\alpha/2}$ then for any $p_0 \in \mathcal{P}(\mathcal{M})$ and for any $t \geq 1/2$ we have*

$$\|p_0\mathrm{P}_t - p_{\mathrm{ref}}\|_{\mathrm{TV}} \leq C^{1/2}\mathrm{e}^{\lambda_1/2}\mathrm{e}^{-\lambda_1 t},$$

where λ_1 is the first non-negative eigenvalue of $-\Delta_{\mathcal{M}}$ in $\mathrm{L}^2(p_{\mathrm{ref}})$ and we recall that $(\mathrm{P}_t)_{t\geq 0}$ is the semi-group of the Brownian motion.

A review on lower bounds on the first positive eigenvalue of the Laplace-Beltrami operator can be found in (He, 2013). These lower bounds usually depend on the Ricci curvature of the manifold or its diameter. We conclude this section by noting that in the non-compact case (Li, 1986) establishes similar estimates in the case of a manifold with non-negative Ricci curvature and maximal volume growth.

S3 Difference between ODE and SDE likelihood computations

In this section, we show that the likelihood computation from Song et al. (2021) does not coincide with the likelihood computation obtained with the SDE model. We present our findings in the Riemannian setting but our conclusions can be adapted to the Euclidean setting with arbitrary forward dynamics. Recall that we consider a Brownian motion on the manifold as a forward process $(\mathbf{B}_t^{\mathcal{M}})_{t\in[0,T]}$ with $\{p_t\}_{t=0}^T$ the associated family of densities. We have that for any $t \in [0,T]$ and $x \in \mathcal{M}$

$$\partial_t p_t(x) = \tfrac{1}{2}\Delta p_t(x) = \operatorname{div}(\tfrac{1}{2} p_t \nabla \log p_t)(x). \tag{S3}$$

ODE model In the case of the ODE model we define $(\mathbf{X}_t)_{t\in[0,T]}$ such that $\mathbf{X}_0$ has distribution π and satisfies $\mathrm{d}\mathbf{X}_t = \frac{1}{2}\nabla \log p_t(\mathbf{X}_t)\mathrm{d}t$. Note that the family of densities $\{q_t\}_{t=0}^T$ associated with $(\mathbf{X}_t)_{t\in[0,T]}$ also satisfies Equation (S3). Now, we consider $(\hat{\mathbf{X}}_t)_{t\in[0,T]} = (\mathbf{X}_{T-t})_{t\in[0,T]}$ and note that it satisfies

$$\mathrm{d}\hat{\mathbf{X}}_t = -\tfrac{1}{2}\nabla \log p_{T-t}(\hat{\mathbf{X}}_t)\mathrm{d}t. \tag{S4}$$

Finally, we consider $(\mathbf{Y}_t^{\mathrm{ODE}})_{t\in[0,T]}$ which also satisfies Equation (S4) and such that the distribution of $\mathbf{Y}_0^{\mathrm{ODE}}$ is p_{ref}. Denoting $\{q_t^{\mathrm{ODE}}\}_{t=0}^T$ the densities of $(\mathbf{Y}_t^{\mathrm{ODE}})_{t\in[0,T]}$ w.r.t. p_{ref} we have for any $t \in [0,T]$ and $x \in \mathcal{M}$

$$\partial_t q_t^{\mathrm{ODE}}(x) = \operatorname{div}(q_t^{\mathrm{ODE}} - \tfrac{1}{2}\nabla \log p_{T-t})(x). \tag{S5}$$

SDE model When sampling we consider a process $(\mathbf{Y}_t^{\mathrm{SDE}})_{t\in[0,T]}$ such that $\mathbf{Y}_0^{\mathrm{SDE}}$ has distribution p_{ref} and whose family of densities $\{q_t^{\mathrm{SDE}}\}_{t=0}^T$ satisfies for any $t \in [0,T]$ and $x \in \mathcal{M}$

$$\partial_t q_t^{\mathrm{SDE}}(x) = -\operatorname{div}(\log p_{T-t} q_t^{\mathrm{SDE}}(x)) + \tfrac{1}{2}\Delta q_t^{\mathrm{SDE}}(x) = \operatorname{div}(q_t^{\mathrm{SDE}}\{-\nabla \log p_{T-t} + \tfrac{1}{2}\nabla \log q_t^{\mathrm{SDE}}\})(x). \tag{S6}$$

Hence, Equation (S5) and Equation (S6) do not agree, except if $q_t^{\mathrm{SDE}} = q_t^{\mathrm{ODE}} = p_{T-t}$ which is the case if and only if $\mathbf{Y}_0^{\mathrm{SDE}}$ and $\mathbf{Y}_0^{\mathrm{ODE}}$ have the same distribution as $\mathbf{X}_T$. Note that it is possible to evaluate the likelihood of the SDE model using that

$$\partial_t \log q_t^{\mathrm{SDE}}(\mathbf{Y}_t^{\mathrm{SDE}}) = \operatorname{div}(-\nabla \log p_{T-t}(\mathbf{Y}_t^{\mathrm{SDE}}) + \tfrac{1}{2}\nabla \log q_t^{\mathrm{SDE}}(\mathbf{Y}_t^{\mathrm{SDE}}))\mathrm{d}t.$$

We can use the score approximation $\boldsymbol{s}_\theta(t,x)$ to approximate $\nabla \log p_t(x)$ for any $t \in [0,T]$ and $x \in \mathcal{M}$. In order to approximate $\nabla \log q_t^{\mathrm{SDE}}$, one can consider another neural network $\boldsymbol{t}_\theta(t,x)$ approximating $\nabla \log q_t^{\mathrm{SDE}}(x)$ for any $t \in [0,T]$ and $x \in \mathcal{M}$. This approximation can be obtained using the implicit score loss presented in Section 4.3.

S4 Eigenfunctions, eigenvalues of the Laplace-Beltrami operator

In this section, we recall the eigenfunctions and eigenvalues of the Laplace-Beltrami operator in two specific cases: the d-dimensional torus and the d-dimensional sphere.

The case of the torus Let $\{b_i\}_{i=1}^d$ be a basis of $\mathbb{R}^d$. We consider the associated lattice on $\mathbb{R}^d$, i.e. $\Gamma = \{\sum_{i=1}^d \alpha_i b_i \ : \ \{\alpha_i\}_{i=1}^d \in \mathbb{Z}^d\}$. Finally, the associated d-dimensional torus is defined as $\mathbb{T}_\Gamma = \mathbb{R}^d/\Gamma$. Denote $\mathrm{B} = (b_1, \dots, b_d) \in \mathbb{R}^{d\times d}$. Let $\{\bar{b}_i\}_{i=1}^d \in (\mathbb{R}^d)^d$ such that $(\mathrm{B}^{-1})^\top = (\bar{b}_1, \dots, \bar{b}_d)$. We define $\Gamma^\star = \{\sum_{i=1}^d \alpha_i \bar{b}_i \ : \ \{\alpha_i\}_{i=1}^d \in \mathbb{Z}^d\}$, the dual lattice. Note that for any $x \in \Gamma$ and $y \in \Gamma^\star$ we have that $\langle x, y\rangle \in \mathbb{Z}$ and that if $\{b_i\}_{i=1}^d$ is an orthonormal basis then $\Gamma = \Gamma^\star$. The torus $\mathbb{R}^d/\Gamma$ is a (flat) compact Riemannian manifold. The set of eigenvalues of the Laplace-Beltrami operator is given by $\{-4\pi^2\|y\|^2 \ : \ y \in \Gamma^\star\}$. The eigenfunctions of the Laplace-Beltrami operator are given by $\{x \mapsto \sin(2\pi\langle x, y\rangle) \ : \ y \in \Gamma^\star\}$ and $\{x \mapsto \cos(2\pi\langle x, y\rangle) \ : \ y \in \Gamma^\star\}$.

The case of the sphere Next, we investigate the case of the d-dimensional sphere (see Saloff-Coste, 1994). The set of eigenvalues of the Laplace-Beltrami operator is given by $\{-k(k+d-1) \ : \ k \in \mathbb{N}\}$. Note that $\lambda_k = k(k+d-1)$ has multiplicity $d_k = (k+d-2)!/\{(d-1)!k\}(2k+d-1)$. The eigenfunctions of the Laplace-Beltrami operator are known as the spherical harmonics and can be defined in terms of Legendre polynomials. When investigating the heat kernel on the d-dimensional sphere, we are interested in the product $(x,y) \mapsto \sum_{\phi\in\Phi_n} \phi(x)\phi(y)$, where Φ_n is the set of eigenfunctions associated with the eigenvalue

λ_n for $n \in \mathbb{N}$. This function can be described using the Gegenbauer polynomials (see Atkinson and Han, 2012, Theorem 2.9). More precisely, we have that for any $n \in \mathbb{N}$ and $x, y \in \mathbb{S}^d$

$$\begin{aligned} G_n(x,y) &= \textstyle\sum_{\phi \in \Phi_n} \phi(x)\phi(y) \\ &= n!\Gamma((d-1)/2) \textstyle\sum_{k=0}^{\lfloor n/2 \rfloor} (-1)^k (1-\langle x,y \rangle^2)\langle x,y \rangle^{n-2k} / (4^k k!(n-2k)!\Gamma(k+(d-1)/2)), \end{aligned}$$

where here $\Gamma : \mathbb{R}_+ \to \mathbb{R}$ is given for any $v > 0$ by $\Gamma(v) = \int_0^{+\infty} t^{v-1} \mathrm{d}t^{-t} \mathrm{d}t$. In the special case where $d = 1$, then the heat kernel coincide with the wrapped Gaussian density and can be easily evaluated.

S5 Time-reversal formula: extension to compact Riemannian manifolds

In this section, we provide the proof of Theorem 5. The proof follows the arguments of Cattiaux et al. (2021, Theorem 4.9). We could have also applied the abstract results of Cattiaux et al. (2021, Theorem 5.7) to obtain our results. Note that the time-reversal on manifold could also be obtained by readily extending arguments from Haussmann and Pardoux (1986), however the entropic conditions found by Cattiaux et al. (2021) are more natural when it comes to the study of the Schrödinger Bridge problem. For the interested reader we provide an informal derivation of the time-reversal formula obtained by Haussmann and Pardoux (1986) in Appendix S5.1. The proof of Theorem 5 is given in Appendix S5.2. Finally, we emphasize that García-Zelada and Huguet (2021) develops a Girsanov theory for stochastic processes defined on compact manifolds with boundary in order to study the Brenier-Schrödinger problem.

S5.1 Informal derivation

In this section, we provide a non-rigorous derivation of Theorem 5 following the approach of Haussmann and Pardoux (1986). Let $(\mathbf{X}_t)_{t\in[0,T]}$ be a continuous process such that for any $f \in \mathrm{C}^2(\mathcal{M})$ we have that $(\mathbf{M}_t^{\mathbf{X},f})_{t\in[0,T]}$ is a $\mathbf{X}$-martingale where for any $t \in [0,T]$

$$\mathbf{M}_t^{\mathbf{X},f} = f(\mathbf{X}_t) - \textstyle\int_0^t \{\langle b(\mathbf{X}_s), \nabla f(\mathbf{X}_s)\rangle_{\mathcal{M}} + (1/2)\Delta f(\mathbf{X}_s)\}\mathrm{d}s. \tag{S7}$$

Let $(\mathbf{Y}_t)_{t\in[0,T]} = (\mathbf{X}_{T-t})_{t\in[0,T]}$. Our goal is to show that for any $f \in \mathrm{C}^2(\mathcal{M})$, $(\mathbf{M}_t^{\mathbf{Y},f})_{t\in[0,T]}$ is a $\mathbf{Y}$-martingale where for any $t \in [0,T]$

$$\mathbf{M}_t^{\mathbf{Y},f} = f(\mathbf{Y}_t) - \textstyle\int_0^t \{\langle b(\mathbf{Y}_s) + \nabla \log p_{T-s}(\mathbf{Y}_s), \nabla f(\mathbf{Y}_s)\rangle_{\mathcal{M}} + (1/2)\Delta f(\mathbf{Y}_s)\}\mathrm{d}s.$$

Note that here we implicitly assume that for any $t \in [0,T]$, $\mathbf{X}_t$ admits a smooth positive density w.r.t. p_{ref} denoted p_t. In other words, we want to show that for any $g \in \mathrm{C}^2(\mathcal{M})$ and $s, t \in [0,T]$ with $t \geq s$ we have

$$\mathbb{E}[g(\mathbf{Y}_s)(f(\mathbf{Y}_t) - f(\mathbf{Y}_s))] = \mathbb{E}[g(\mathbf{Y}_s)\textstyle\int_s^t \{\langle b(\mathbf{Y}_u) + \nabla \log p_{T-u}(\mathbf{Y}_u), \nabla f(\mathbf{Y}_u)\rangle_{\mathcal{M}} + (1/2)\Delta f(\mathbf{Y}_u)\}\mathrm{d}u]. \tag{S8}$$

We introduce the infinitesimal generator $\mathcal{A} : \mathrm{C}^2(\mathcal{M}) \to \mathrm{C}(\mathcal{M})$ given for any $f \in \mathrm{C}^2(\mathcal{M})$ and $x \in \mathcal{M}$ by

$$\mathcal{A}(f)(x) = \langle b(x), \nabla f(x)\rangle_{\mathcal{M}} + (1/2)\Delta f(x).$$

Similarly, we introduce the infinitesimal generator $\tilde{\mathcal{A}} : [0,T] \times \mathrm{C}^2(\mathcal{M}) \to \mathrm{C}(\mathcal{M})$ given for any $f \in \mathrm{C}^2(\mathcal{M})$, $t \in [0,T]$ and $x \in \mathcal{M}$ by

$$\tilde{\mathcal{A}}(t,f)(x) = \langle b(x) + \nabla \log p_{T-t}(x), \nabla f(x)\rangle_{\mathcal{M}} + (1/2)\Delta f(x).$$

With these notations, (S9) can be written as follows: we want to show that for any $g \in \mathrm{C}^2(\mathcal{M})$ and $s, t \in [0,T]$ with $t \geq s$ we have

$$\mathbb{E}[g(\mathbf{Y}_s)(f(\mathbf{Y}_t) - f(\mathbf{Y}_s))] = \mathbb{E}[g(\mathbf{Y}_s)\textstyle\int_s^t \tilde{\mathcal{A}}(u, \mathbf{Y}_u)\mathrm{d}u]. \tag{S9}$$

The rest of this section follows the first part of the proof of Haussmann and Pardoux (1986, Theorem 2.1). Let $t, s \in [0,T]$ with $t \geq s$. We have

$$\begin{aligned} \mathbb{E}[g(\mathbf{Y}_s)(f(\mathbf{Y}_t) - f(\mathbf{Y}_s))] &= \mathbb{E}[g(\mathbf{X}_{T-s})(f(\mathbf{X}_{T-t}) - f(\mathbf{X}_{T-t}))] \\ &= \mathbb{E}[\mathbb{E}[g(\mathbf{X}_{T-s})|\mathbf{X}_{T-t}]f(\mathbf{X}_{T-t})] - \mathbb{E}[g(\mathbf{X}_{T-s})f(\mathbf{X}_{T-s})] \end{aligned}$$

$$= \mathbb{E}[v(T-t, \mathbf{X}_{T-t}) f(\mathbf{X}_{T-t})] - \mathbb{E}[v(T-s, \mathbf{X}_{T-s}) f(\mathbf{X}_{T-s})], \quad \text{(S10)}$$

with $v : [0, T-s] \times \mathcal{M} \to \mathbb{R}$ given for any $u \in [0, T-s]$ and $x \in \mathcal{M}$ by $v(u, x) = \mathbb{E}[g(\mathbf{X}_{T-s}) | \mathbf{X}_u = x]$. We have that v satisfies the backward Kolmogorov equation, i.e. we have for any $u \in [0, T-s]$ and $x \in \mathcal{M}$

$$\partial_u v(u, x) = -\mathcal{A} v(u, x). \quad \text{(S11)}$$

Note that it is not trivial to show that v is regular enough to satisfy the backward Kolmogorov equation. In this informal derivation, we assume that v is regular enough and will provide a different rigorous proof of the time-reversal formula in Appendix S5.2. However, note that it is possible to show that v indeed satisfies the backward Kolmogorov equation by adapting arguments from Haussmann and Pardoux (1986) to the manifold framework.

Let $h : [0, T-s] \times \mathcal{M} \to \mathbb{R}$ given for any $u \in [0, T-s]$ and $x \in \mathcal{M}$ by $h(u, x) = v(u, x) f(x)$. Using (S11), we have for any $u \in [0, T-s]$ and $x \in \mathcal{M}$

$$\begin{aligned} \partial_u h(u, x) + \mathcal{A} h(u, x) &= f(x) \partial_u v(u, x) + f(x) \mathcal{A} v(u, x) + v(u, x) \mathcal{A} f(x) + \langle \nabla f(x), \nabla v(u, x) \rangle \quad \text{(S12)} \\ &= v(u, x) \mathcal{A} f(x) + \langle \nabla f(x), \nabla v(u, x) \rangle_{\mathcal{M}}. \end{aligned}$$

In addition, using the divergence theorem (see Lee, 2018, p.51), we have for any $u \in [0, T-s]$

$$\begin{aligned} \mathbb{E}[\langle \nabla f(\mathbf{X}_u), \nabla v(u, \mathbf{X}_u) \rangle_{\mathcal{M}}] &= \textstyle\int_{\mathcal{M}} \langle \nabla f(x_u), \nabla v(u, x_u) p_u(x_u) \rangle_{\mathcal{M}} \mathrm{d}p_{\mathrm{ref}}(x_u) \\ &= -\textstyle\int_{\mathcal{M}} v(u, x_u) \mathrm{div}(p_u \nabla f)(x_u) \mathrm{d}p_{\mathrm{ref}}(x_u) \\ &= -\textstyle\int_{\mathcal{M}} v(u, x_u) \Delta f(x_u) p_u(x_u) \mathrm{d}p_{\mathrm{ref}}(x_u) - \textstyle\int_{\mathcal{M}} v(u, x_u) \langle \nabla f(x_u), \nabla \log p_u(x_u) \rangle_{\mathcal{M}} p_u(x_u) \mathrm{d}p_{\mathrm{ref}}(x_u) \\ &= -\mathbb{E}[v(u, \mathbf{X}_u) \Delta f(\mathbf{X}_u)] - \mathbb{E}[v(u, \mathbf{X}_u) \langle \nabla f(\mathbf{X}_u), \nabla \log p_u(\mathbf{X}_u) \rangle_{\mathcal{M}}]. \end{aligned}$$

Therefore, using this result and (S12) we get that for any $u \in [0, T-s]$

$$\begin{aligned} \mathbb{E}[\partial_u h(u, \mathbf{X}_u) + \mathcal{A} h(u, \mathbf{X}_u)] &= \mathbb{E}[v(u, \mathbf{X}_u) \{ \langle b(\mathbf{X}_u) - \nabla \log p_u(\mathbf{X}_u), \nabla f(\mathbf{X}_u) \rangle_{\mathcal{M}} - (1/2) \Delta f(\mathbf{X}_u) \}] \\ &= -\mathbb{E}[v(u, \mathbf{X}_u) \tilde{\mathcal{A}}(T-u, f)(\mathbf{X}_u)]. \end{aligned}$$

Combining this result and (S7) and that for any $u \in [0, T-s]$ and $x \in \mathcal{M}$, $v(u, x) = \mathbb{E}[g(\mathbf{X}_{T-s}) | \mathbf{X}_u = x]$ we get

$$\begin{aligned} \mathbb{E}[v(T-t, \mathbf{X}_{T-t}) f(\mathbf{X}_{T-t})] - \mathbb{E}[v(T-s, \mathbf{X}_{T-s}) f(\mathbf{X}_{T-s})] &= \mathbb{E}[h(T-t, \mathbf{X}_{T-t}) - h(T-s, \mathbf{X}_{T-s})] \\ &= \textstyle\int_{T-t}^{T-s} \mathbb{E}[v(u, \mathbf{X}_u) \tilde{\mathcal{A}}(T-u, \mathbf{X}_u)] \mathrm{d}u \\ &= \mathbb{E}[g(\mathbf{X}_{T-s}) \textstyle\int_{T-t}^{T-s} \tilde{\mathcal{A}}(T-u, \mathbf{X}_u) \mathrm{d}u]. \end{aligned}$$

Using this result, (S10) and the change of variable $u \mapsto T-u$ we obtain

$$\mathbb{E}[g(\mathbf{Y}_s)(f(\mathbf{Y}_t) - f(\mathbf{Y}_s))] = \mathbb{E}[g(\mathbf{X}_{T-s}) \textstyle\int_{T-t}^{T-s} \tilde{\mathcal{A}}(u, \mathbf{X}_u) \mathrm{d}u] = \mathbb{E}[g(\mathbf{Y}_s) \textstyle\int_s^t \tilde{\mathcal{A}}(u, \mathbf{Y}_u) \mathrm{d}u].$$

Hence, (S9) holds and we have proved Theorem 5. Again, we emphasize that in order to make the proof completely rigourous one needs to derive regularity properties of v.

S5.2 Proof of Theorem 5

In this section, we follow another approach to prove the time-reversal formula. We are going to use the integration by part formula of Cattiaux et al. (2021, Theorem 3.17) in a similar spirit as Cattiaux et al. (2021, Theorem 4.9) in the Euclidean setting. In order to adapt arguments from Cattiaux et al. (2021) to our Riemannian setting, we use the Nash embedding theorem in order to embed our processes in a Euclidean space and leverage tools from Girsanov theory. The rest of the section is organized as follows. First in Appendix S5.2.1, we recall basic properties of infinitesimal generators and recall the integration by part formula of Cattiaux et al. (2021, Theorem 3.17). Then in Appendix S5.2.2, we extend some Girsanov theory to compact Riemannian manifolds using the Nash embedding theorem. We conclude the proof in Appendix S5.2.3.

S5.2.1 Diffusion processes and integration by part formula

In this section, we state a simplified version of Cattiaux et al. (2021, Theorem 3.17) for Markov continuous path (probability) measure on Polish spaces. Let $(\mathsf{X}, \mathcal{X})$ be a Polish space. We say that $\mathbb{P}$ is a path measure if $\mathbb{P} \in \mathcal{P}(\mathrm{C}([0,T], \mathsf{X}))$. Let $(\mathbf{X}_t)_{t\in[0,T]}$ with distribution $\mathbb{P}$. We denote $(\mathcal{F}_t)_{t\in[0,T]}$ the filtration such that for any $t \in [0,T]$, $\mathcal{F}_t = \sigma(\mathbf{X}_s,\ s \in [0,t])$. Let $(\mathbf{M}_t)_{t\in[0,T]}$ be a Polish-valued stochastic process. We say that $(\mathbf{M}_t)_{t\in[0,T]}$ is a $\mathbb{P}$-local martingale if it is a local martingale w.r.t. the filtration $(\mathcal{F}_t)_{t\in[0,T]}$. A function $u : [0,T] \times \mathsf{X} \to \mathbb{R}$ is said to be in the domain of the extended generator of $\mathbb{P}$ if there exists a process $(\bar{\mathcal{A}}_{\mathbb{P}} u(t, \mathbf{X}_{[0,t]}))_{t\in[0,T]}$ such that:

(a) $(\bar{\mathcal{A}}_{\mathbb{P}} u(t, \mathbf{X}_{[0,t]}))_{t\in[0,T]}$ is adapted w.r.t. $(\mathcal{F}_t)_{t\in[0,T]}$.

(b) $\int_0^T |\bar{\mathcal{A}}_{\mathbb{P}} u(t, \mathbf{X}_{[0,t]})| \mathrm{d}t < +\infty$, $\mathbb{P}$-a.s.

(c) The process $(\mathbf{M}_t)_{t\in[0,T]}$ is a $\mathbb{P}$-local martingale, where for any $t \in [0,T]$

$$\mathbf{M}_t = u(t, \mathbf{X}_t) - u(0, \mathbf{X}_0) - \int_0^t \bar{\mathcal{A}}_{\mathbb{P}} u(s, \mathbf{X}_{[0,s]}) \mathrm{d}s.$$

The domain of the extended generator is denoted $\mathrm{dom}(\bar{\mathcal{A}}_{\mathbb{P}})$. We say that (u,v) with $u, v : [0,T] \times \mathsf{X} \to \mathbb{R}$ is in the domain of the carré du champ if $u, v, uv \in \mathrm{dom}(\bar{\mathcal{A}}_{\mathbb{P}})$. In this case, we define the carré du champ $\bar{\Upsilon}_{\mathbb{P}}$ as

$$\bar{\Upsilon}_{\mathbb{P}}(u,v) = \bar{\mathcal{A}}_{\mathbb{P}}(uv) - \bar{\mathcal{A}}_{\mathbb{P}}(u)v - \bar{\mathcal{A}}_{\mathbb{P}}(v)u.$$

Note that if $\mathsf{X} = \mathcal{M}$ is a Riemannian manifold, $\mathrm{C}^2(\mathcal{M}) \subset \mathrm{dom}(\bar{\mathcal{A}}_{\mathbb{P}})$ and for any $u \in \mathrm{C}^2(\mathcal{M})$ $\bar{\mathcal{A}}_{\mathbb{P}}(u) = \langle \nabla u, X \rangle_{\mathcal{M}} + (1/2)\Delta u$ with $X \in \Gamma(\mathrm{T}\mathcal{M})$ then we have that $\mathrm{C}^2(\mathcal{M}) \times \mathrm{C}^2(\mathcal{M}) \subset \mathrm{dom}(\bar{\Upsilon}_{\mathbb{P}})$ and for any $u, v \in \mathrm{C}^2(\mathcal{M})$, $\bar{\Upsilon}_{\mathbb{P}}(u,v) = \langle \nabla u, \nabla v \rangle_{\mathcal{M}}$. Assume that there exists $\mathcal{U}_{\mathbb{P}} \subset \mathrm{dom}(\bar{\mathcal{A}}_{\mathbb{P}}) \cap \mathrm{C}_b(\mathsf{X})$ such that $\mathcal{U}_{\mathbb{P}}$ is an algebra. We denote $\mathcal{U}_{\mathbb{P},2}$ such that

$$\mathcal{U}_{\mathbb{P},2} = \{u \in \mathcal{U}_{\mathbb{P}} \ : \ \bar{\mathcal{A}}_{\mathbb{P}} u \in \mathrm{L}^2(\mathbb{P}),\ \bar{\Upsilon}_{\mathbb{P}}(u,u) \in \mathrm{L}^1(\mathbb{P})\}.$$

Finally we denote $R(\mathbb{P})$ the time-reverse path measure, i.e. for any $\mathsf{A} \in \mathcal{B}(\mathrm{C}([0,T], \mathsf{X}))$ we have $R(\mathbb{P})(\mathsf{A}) = \mathbb{P}(R(\mathsf{A}))$, where $R(\mathsf{A}) = \{t \mapsto \omega_{T-t} \ : \ \omega \in \mathsf{A}\}$. In what follows, we assume $\mathbb{P}$ is Markov. It is well-known, see (Léonard et al., 2014, Theorem 1.2) for instance, that in this case $R(\mathbb{P})$ is also Markov. In addition, since $\mathbb{P}$ is Markov, for any $u \in \mathrm{dom}(\bar{\mathcal{A}}_{\mathbb{P}})$ and $t \in [0,T]$ there exists $\mathcal{A}_{\mathbb{P}}$ such that $\bar{\mathcal{A}}_{\mathbb{P}} u(t, \mathbf{X}_{[0,t]}) = \mathcal{A}_{\mathbb{P}} u(t, \mathbf{X}_t)$ with $\mathcal{A}_{\mathbb{P}} u : [0,T] \times \mathsf{X} \to \mathbb{R}$. Similarly, we define $\Upsilon_{\mathbb{P}}(u,v) : [0,T] \times \mathsf{X} \to \mathbb{R}$ from $\bar{\Upsilon}_{\mathbb{P}}(u,v)$.

We are now ready to state the integration by part formula, (Cattiaux et al., 2021, Theorem 3.17).

Theorem S16. *Let $u, v \in \mathcal{U}_{\mathbb{P},2}$. The following hold:*

(a) If $u \in \mathrm{dom}(\mathcal{A}_{R(\mathbb{P})})$ and $\mathcal{A}_{R(\mathbb{P})} u \in \mathrm{L}^1(\mathbb{P})$ then for almost any $t \in [0,T]$

$$\mathbb{E}[\{\mathcal{A}_{\mathbb{P}} u(t, \mathbf{X}_t) + \mathcal{A}_{R(\mathbb{P})} u(T-t, \mathbf{X}_t)\} v(\mathbf{X}_t) + \Upsilon_{\mathbb{P}}(u,v)(t, \mathbf{X}_t)] = 0.$$

(b) If the following hold:

i) $\Upsilon_{\mathbb{P}}(u,v) \in \mathrm{C}([0,T] \times \mathsf{X}, \mathbb{R})$.

ii) $\mathcal{U}_{2,\mathbb{P}}$ determines the weak convergence of Borel measure.

iii) μ defines a finite measure on $[0,T] \times \mathsf{X}$ where for any $\omega \in \bar{\mathcal{U}}_{2,\mathbb{P}}$ we have

$$\mu[\omega] = \mathbb{E}[\int_0^T \Upsilon_{\mathbb{P}}(u, \omega_t)(t, \mathbf{X}_t) \mathrm{d}t,]$$

where $\bar{\mathcal{U}}_{2,\mathbb{P}} = \{\omega \in \mathrm{C}([0,T] \times \mathsf{X}, \mathbb{R}) \ : \ \omega(t, \cdot) \in \mathcal{U}_{2,\mathbb{P}} \ \text{ for any } t \in [0,T]\}$.

Then $u \in \mathrm{dom}(\mathcal{A}_{R(\mathbb{P})})$ and $\mathcal{A}_{R(\mathbb{P})} u \in \mathrm{L}^1(\mathbb{P})$.

Note that this theorem is a simplified version of Cattiaux et al. (2021, Theorem 3.17) where we restrict ourselves to the case of Markov path measures. In what follows, we wish to apply Theorem S16 to diffusion processes on manifolds. To do so, we will verify that under a finite entropy assumption, the conditions $u \in \mathrm{dom}(\mathcal{A}_{R(\mathbb{P})})$ and $\mathcal{A}_{R(\mathbb{P})} u \in \mathrm{L}^1(\mathbb{P})$ are fullfilled for a class of regular functions u. These integrability results are obtained using Girsanov theory.

S5.2.2 Girsanov theory on compact Riemannian manifolds

In this section, we will consider two types of martingale problems: one on Euclidean spaces and one on the compact Riemannian manifold $\mathcal{M}$. Let $\mathbb{P} \in \mathcal{P}(\mathrm{C}([0,T], \mathbb{R}^p))$. We say that $\mathbb{P}$ satisfies the (Euclidean) martingale problem with infinitesimal generator $\mathcal{A}: [0,T] \times \mathrm{C}^2(\mathbb{R}^p) \times \mathbb{R}^p \to \mathbb{R}$ if for any $u \in \mathrm{C}_c^2(\mathbb{R}^p)$, $(\mathbf{M}_t)_{t\in[0,T]}$ is a $\mathbb{P}$-martingale where for any $t \in [0,T]$ we have

$$\mathbf{M}_t = \mathbf{M}_0 + \int_0^t \mathcal{A}(t,u)(\mathbf{X}_s)\mathrm{d}s,$$

where $(\mathbf{X}_t)_{t\in[0,T]}$ has distribution $\mathbb{P}$ and $\int_0^T |\mathcal{A}(t,u)(\mathbf{X}_s)\mathrm{d}t| < +\infty$, $\mathbb{P}$-a.s. Let $\mathbb{P} \in \mathcal{P}(\mathrm{C}([0,T], \mathcal{M}))$. We say that $\mathbb{P}$ satisfies the (Riemannian) martingale problem with infinitesimal generator $\tilde{\mathcal{A}}: [0,T] \times \mathrm{C}^2(\mathcal{M}) \times \mathcal{M} \to \mathbb{R}$ if for any $u \in \mathrm{C}^2(\mathcal{M})$, $(\mathbf{M}_t)_{t\in[0,T]}$ is a $\mathbb{P}$-martingale where for any $t \in [0,T]$ we have

$$\mathbf{M}_t = \mathbf{M}_0 + \int_0^t \tilde{\mathcal{A}}(t,u)(\mathbf{X}_s)\mathrm{d}s,$$

where $(\mathbf{X}_t)_{t\in[0,T]}$ has distribution $\mathbb{P}$ and $\int_0^T |\tilde{\mathcal{A}}(t,u)(\mathbf{X}_s)\mathrm{d}t| < +\infty$, $\mathbb{P}$-a.s. We now prove the following theorem.

Proposition S17. *Let $\mathbb{Q}$ be the path measure of a Brownian motion on $\mathcal{M}$. Let $\mathbb{P}$ be a Markov path measure on $\mathrm{C}([0,T], \mathcal{M})$ such that $\mathrm{KL}(\mathbb{P}|\mathbb{Q}) < +\infty$. Then there exists β such that for any $t \in [0,T]$ and $x \in \mathcal{M}$, $\beta(t,x) \in \mathrm{T}_x\mathcal{M}$. In addition, we have that $\mathbb{P}$ satisfies the martingale problem with infinitesimal generator $\mathcal{A}$ where for any $t \in [0,T]$, $u \in \mathrm{C}^2(\mathcal{M})$ and $x \in \mathcal{M}$ we have*

$$\mathcal{A}(t,u)(x) = \langle \beta(t,x), \nabla u(x)\rangle_{\mathcal{M}} + (1/2)\Delta u(x).$$

In addition, we have that

$$\mathrm{KL}(\mathbb{P}|\mathbb{Q}) = \mathrm{KL}(\mathbb{P}_0|\mathbb{Q}_0) + (1/2)\int_0^T \mathbb{E}[\|\beta(t,\mathbf{X}_t)\|^2]\mathrm{d}t,$$

where $(\mathbf{X}_t)_{t\in[0,T]}$ has distribution $\mathbb{P}$.

Proof. First, we extend $(\mathbf{B}_t^{\mathcal{M}})_{t\in[0,T]}$ to $\mathbb{R}^p$ using the Nash embedding theorem (see Gunther, 1991). $(\mathbf{B}_t^{\mathcal{M}})_{t\in[0,T]}$ can be seen as a process on $\mathbb{R}^p$ (for some $p \in \mathbb{N}$) which satisfies in a weak sense

$$\mathrm{d}\mathbf{B}_t^{\mathcal{M}} = \sum_{i=1}^p P_i(\mathbf{B}_t^{\mathcal{M}}) \circ \mathrm{d}\mathbf{B}_t^i = P(\mathbf{B}_t^{\mathcal{M}}) \circ \mathrm{d}\mathbf{B}_t,$$

where $(\mathbf{B}_t)_{t\in[0,T]}$ is a p-dimensional Brownian motion and $P \in \mathrm{C}^\infty(\mathbb{R}^p, \mathbb{R}^{p\times p})$ is such that for any $x \in \mathcal{M}$, $P(x)$ is the projection onto $\mathrm{T}_x\mathcal{M}$ and for any $i \in \{1,\dots,p\}$, $P_i \in \mathrm{C}^\infty(\mathbb{R}^p, \mathbb{R}^p)$ with $P_i = Pe_i$ where $\{e_j\}_{j=1}^d$ is the canonical basis of $\mathbb{R}^p$. Using the link between Stratanovitch and Itô integral, there exists $\bar{b} \in \mathrm{C}^\infty(\mathbb{R}^p, \mathbb{R}^p)$ such that $(\mathbf{B}_t^{\mathcal{M}})_{t\in[0,T]}$ can be seen as a process on $\mathbb{R}^p$ which satisfies in a weak sense

$$\mathrm{d}\mathbf{B}_t^{\mathcal{M}} = \bar{b}(\mathbf{B}_t^{\mathcal{M}})\mathrm{d}t + P(\mathbf{B}_t^{\mathcal{M}})\mathrm{d}\mathbf{B}_t.$$

For any $u \in \mathrm{C}^2(\mathcal{M})$, we consider $\bar{u}$ an extension to $\mathrm{C}_c^2(\mathbb{R}^p)$ and we have for any $s,t \in [0,T]$

$$\begin{aligned}&\mathbb{E}[\bar{v}(\mathbf{B}_s^{\mathcal{M}})\int_s^t (1/2)\Delta u(\mathbf{B}_u^{\mathcal{M}})\mathrm{d}u] \\ &\qquad = \mathbb{E}[\bar{v}(\mathbf{B}_s^{\mathcal{M}})\int_s^t \{\langle \nabla\bar{u}(\mathbf{B}_u^{\mathcal{M}}), \bar{b}(\mathbf{B}_u^{\mathcal{M}})\rangle + (1/2)\langle P(\mathbf{B}_u^{\mathcal{M}}), \nabla^2\bar{u}(\mathbf{B}_u^{\mathcal{M}})\rangle\}\mathrm{d}u].\end{aligned}$$

In particular, we get that for any $x \in \mathcal{M}$, $\Delta u(x) = 2\langle \bar{u}(x), \bar{b}(x)\rangle + \Delta\bar{u}(x)$. Note that $(\mathbf{B}_t^{\mathcal{M}})_{t\in[0,T]}$ (seen as a process on $\mathbb{R}^p$) satisfies the condition (U) in Léonard (2012b). Therefore applying (Léonard, 2012b, Theorem 2.1), (Cattiaux et al., 2021, Claim 4.5), there exists $\bar{\beta}: [0,T] \times \mathbb{R}^p \to \mathbb{R}^p$ such that

$$\mathrm{KL}(\mathbb{P}|\mathbb{Q}) = \mathrm{KL}(\mathbb{P}_0|\mathbb{Q}_0) + (1/2)\int_0^T \mathbb{E}[\|P(\mathbf{X}_t)\bar{\beta}(t,\mathbf{X}_t)\|^2]\mathrm{d}t. \tag{S13}$$

VDB: prove that for the projection this is okay

In addition, $\mathbb{P}$ (seen as a process on $\mathbb{R}^p$) satisfies a martingale problem with infinitesimal generator $\bar{\mathcal{A}}: [0,T] \times \mathrm{C}_c^2(\mathbb{R}^p) \times \mathbb{R}^p \to \mathbb{R}$ such that for any $t \in [0,T]$, $\bar{u} \in \mathrm{C}^2(\mathbb{R}^p)$ and $x \in \mathbb{R}^p$

$$\bar{\mathcal{A}}(t,\bar{u})(x) = \langle \bar{b}(x) + P(x)\bar{\beta}(t,x), \nabla\bar{u}(x)\rangle + (1/2)\Delta\bar{u}(x).$$

Let $\beta: [0,T] \times \mathcal{M}$ such that for any $t \in [0,T]$ and $x \in \mathcal{M}$ we have $\beta(t,x) = P(x)\bar{\beta}(t,x)$. In particular, we have that for any $x \in \mathcal{M}$, $\beta(t,x) \in \mathrm{T}_x\mathcal{M}$. Let $u \in \mathrm{C}^2(\mathcal{M})$ and consider an extension $\bar{u}$ to $\mathrm{C}^2(\mathbb{R}^p)$.

VDB: dire que c'est okay pour le delta et

For any $t \in [0,T]$ and $x \in \mathcal{M}$ we have

$$\begin{aligned}\bar{\mathcal{A}}(t,\bar{u})(x) &= \langle \bar{b}(x) + P(x)\bar{\beta}(t,x), \nabla \bar{u}(x)\rangle + (1/2)\Delta \bar{u}(x) \\ &= \langle \beta(t,x), \nabla \bar{u}(x)\rangle + (1/2)\Delta u(x) \\ &= \langle P(x)\beta(t,x), P(x)\nabla \bar{u}(x)\rangle + (1/2)\Delta u(x) = \langle \beta(t,x), \nabla u(x)\rangle + (1/2)\Delta u(x).\end{aligned}$$

In particular, we have that $\mathbb{P}$ (seen as a process on $\mathcal{M}$) satisfies a martingale with infinitesimal generator $\bar{\mathcal{A}}: [0,T] \times \mathrm{C}^2_c(\mathcal{M}) \times \mathcal{M} \to \mathbb{R}$ such that for any $t \in [0,T]$, $u \in \mathrm{C}^2(\mathbb{R}^p)$ and $x \in \mathcal{M}$

$$\mathcal{A}(t,\bar{u})(x) = \langle \beta(t,x), \nabla u(x)\rangle_{\mathcal{M}} + (1/2)\Delta u(x).$$

In addition, rewriting (S14) we have

$$\mathrm{KL}\left(\mathbb{P}|\mathbb{Q}\right) = \mathrm{KL}\left(\mathbb{P}_0|\mathbb{Q}_0\right) + (1/2)\int_0^T \mathbb{E}[\|\beta(t,\mathbf{X}_t)\|^2]\mathrm{d}t, \tag{S14}$$

which concludes the proof. □

Once this proposition is established, we can obtain the following straightforward extension of Cattiaux et al. (2021, Proposition 4.6).

Proposition S18. *Let $\mathbb{Q}$ be a Brownian motion with $\mathbb{Q}_0 = p_{\mathrm{ref}}$ and $\mathbb{P}$ a path measure on $\mathrm{C}([0,T],\mathcal{M})$ such that $\mathrm{KL}\left(\mathbb{P}|\mathbb{Q}\right) < +\infty$. Then, there exist $\beta_{\mathbb{P}}, \beta_{R(\mathbb{P})}: [0,T] \times \mathcal{M} \to$ such that for any $t \in [0,T]$ and $x \in \mathcal{M}$, $\beta_{\mathbb{P}}(t,x), \beta_{R(\mathbb{P})}(t,x) \in \mathrm{T}_x\mathcal{M}$. In addition, we have that $\mathbb{P}$ and $R(\mathbb{P})$ satisfy martingale problems with infinitesimal generator $\mathcal{A}_{\mathbb{P}}$, respectively $\mathcal{A}_{R(\mathbb{P})}$ where for any $t \in [0,T]$, $u \in \mathrm{C}^2(\mathcal{M})$ and $x \in \mathcal{M}$ we have*

$$\begin{aligned}\mathcal{A}_{\mathbb{P}}(t,u)(x) &= \langle \beta_{\mathbb{P}}(t,x), \nabla u(x)\rangle_{\mathcal{M}} + (1/2)\Delta u(x), \\ \mathcal{A}_{R(\mathbb{P})}(t,u)(x) &= \langle \beta_{R(\mathbb{P})}(t,x), \nabla u(x)\rangle_{\mathcal{M}} + (1/2)\Delta u(x).\end{aligned}$$

Finally, we have that

$$\int_0^T \mathbb{E}[\|\beta_{\mathbb{P}}(t,\mathbf{X}_t)\|^2]\mathrm{d}t + \int_0^T \mathbb{E}[\left\|\beta_{R(\mathbb{P})}(t,\mathbf{X}_{T-t})\right\|^2]\mathrm{d}t < +\infty,$$

where $(\mathbf{X}_t)_{t\in[0,T]}$ has distribution $\mathbb{P}$.

Proof. The proof is straightforward upon combining Proposition S17 and the fact that $\mathrm{KL}\left(\mathbb{P}|\mathbb{Q}\right) = \mathrm{KL}\left(R(\mathbb{P})|R(\mathbb{Q})\right) = \mathrm{KL}\left(R(\mathbb{P})|\mathbb{Q}\right) < +\infty$, using that $\mathbb{Q}$ is stationary. □

We conclude this section, with the following application of Theorem S16.

Proposition S19. *For any $u, v \in \mathrm{C}^\infty(\mathcal{M})$, we have that for almost any $t \in [0,T]$*

$$\mathbb{E}[v(\mathbf{X}_t)\langle \beta_{\mathbb{P}}(t,\mathbf{X}_t) + \beta_{R(\mathbb{P})}(T-t,\mathbf{X}_t), \nabla u(\mathbf{X}_t)\rangle_{\mathcal{M}} + \langle \nabla u(\mathbf{X}_t), \nabla v(\mathbf{X}_t)\rangle] = 0. \tag{S15}$$

Proof. Remark that $\mathrm{C}^2(\mathcal{M}) \subset \mathrm{dom}(\Upsilon_{\mathbb{P}})$ and $\mathrm{C}^2(\mathcal{M}) \subset \mathrm{dom}(\Upsilon_{R(\mathbb{P})})$. In addition, we have that for any $u, v \in \mathrm{C}^2(\mathcal{M})$, $\Upsilon_{\mathbb{P}}(u,v) = \Upsilon_{R(\mathbb{P})}(u,v) = \langle u, v\rangle$. Note that by Proposition S18 and Theorem S16 we immediately have that for any $u, v \in \mathrm{C}^\infty(\mathcal{M})$, (S15) holds. □

S5.2.3 Concluding the proof

Using Proposition S19 we can now conclude the proof of Theorem 5. First, remark that we can identify $\beta_{\mathbb{P}} = b$. Let $u, v \in \mathrm{C}^\infty(\mathcal{M})$, we have that

$$\mathbb{E}[v(\mathbf{X}_t)\langle b(\mathbf{X}_t) + \beta_{R(\mathbb{P})}(T-t,\mathbf{X}_t), \nabla u(\mathbf{X}_t)\rangle + \Delta u(\mathbf{X}_t)v(\mathbf{X}_t) + \langle \nabla u(\mathbf{X}_t), \nabla v(\mathbf{X}_t)\rangle] = 0.$$

Using that for any $t \in [0,T]$, $\mathbb{P}_t$ admits a smooth positive density w.r.t. p_{ref} denoted p_t and the divergence theorem, see (Lee, 2018, p.51), we have that for any $t \in [0,T]$,

$$\begin{aligned}&\int_{\mathcal{M}}\{\langle \beta_{R(\mathbb{P})}(T-t,x), \nabla u(x)\rangle + \langle b(x), \nabla u(x)\rangle\}v(x)p_t(x)\mathrm{d}p_{\mathrm{ref}}(x) \\ &\qquad = \int_{\mathcal{M}}\langle \nabla u(x)p_t(x), \nabla v(x)\rangle \mathrm{d}p_{\mathrm{ref}}(x) \\ &\qquad = -\int_{\mathcal{M}}\{\Delta u(x) + \langle \nabla \log p_t(x), \nabla u(x)\rangle\}v(x)p_t(x)\mathrm{d}p_{\mathrm{ref}}(x).\end{aligned}$$

Therefore, we get that for any $t \in [0,T]$ and $x \in \mathcal{M}$, $\langle \beta_{R(\mathbb{P})}(T-t,x), \nabla u(x)\rangle = \langle -b(x) + \nabla \log p_t(x), \nabla u(x)\rangle$, which concludes the proof.

S6 Schrödinger Bridges on Manifolds

For Euclidean SGMs, the generative model is given by an approximation of the time-reversal of the noising dynamics $(\mathbf{X}_t)_{t\in[0,T]}$ while the backward dynamics $(\mathbf{Y}_t)_{t\in[0,T]}$ is initialized with the invariant distribution of the noising dynamics (the uniform distribution p_{ref} in case of RSGM). However, in order for the method to yield good results we need $\mathcal{L}(\mathbf{Y}_0) \approx \mathcal{L}(\mathbf{X}_T)$ (see De Bortoli et al., 2021, Theorem 1). Usually, this requires the number of steps in the backward process to be large in order to keep T large and γ small (where $\gamma > 0$ is the stepsize in the Geodesic Random Walk). Another limitation of SGMs is that existing methods target an easy-to-sample reference distribution. Hence, classical SGMs cannot interpolate between two distributions defined by datasets. To circumvent this problem, one can consider a process whose initial and terminal distribution are pinned down using Schrödinger bridges (Schrödinger, 1932; Léonard, 2012a; Chen et al., 2016; De Bortoli et al., 2021).

Dynamical Schrödinger bridges We briefly recall the notion of dynamical Schrödinger bridge (Léonard, 2012a; Chen et al., 2016; Vargas et al., 2021; De Bortoli et al., 2021; Chen et al., 2021a). We consider a reference path probability measure $\mathbb{P} \in \mathcal{P}(\mathrm{C}([0,T], \mathcal{M}))$. In practice, we set $\mathbb{P}$ to be the distribution of the Brownian motion $(\mathbf{B}_t^{\mathcal{M}})_{t\in[0,T]}$ such that $\mathbf{B}_0^{\mathcal{M}}$ has distribution p_0, the target data distribution. Then, we consider the *dynamical Schrödinger bridge problem*

$$\mathbb{Q}^\star = \arg\min\{\mathrm{KL}(\mathbb{Q}|\mathbb{P}) \,:\, \mathbb{Q} \in \mathcal{P}([0,T], \mathcal{M}),\ \mathbb{Q}_0 = p_0,\ \mathbb{Q}_T = p_{\mathrm{ref}}\}.$$

The solution $\mathbb{Q}^\star$ is called the Schrödinger Bridge (SB). Note that if $\mathbb{Q}^\star$ is associated with a backward process $(\mathbf{Y}_t^\star)_{t\in[0,T]}$, then we can obtain a generative model as follows. First sample from $p_{\mathrm{ref}} = \mathcal{L}(\mathbf{Y}_T^\star)$ and then follow the (backward) dynamics of $(\mathbf{Y}_t^\star)_{t\in[0,T]}$. By definition, we obtain that $\mathcal{L}(\mathbf{Y}_0^\star) = p_0$, the target distribution.

In practice however, the solution of the SB problem is approximated using the Iterative Proportional Fitting (IPF) algorithm. Note that in discrete space the IPF is also known as the Sinkhorn algorithm (Sinkhorn, 1967; Peyré and Cuturi, 2019). The IPF defines a sequence of path probability measures $(\mathbb{Q}^n)_{n\in\mathbb{N}} \in (\mathcal{P}(\mathrm{C}([0,T], \mathcal{M})))^{\mathbb{N}}$, such that $\mathbb{Q}^0 = \mathbb{P}$ and for any $n \in \mathbb{N}$

$$\mathbb{Q}^{2n+1} = \arg\min\{\mathrm{KL}\left(\mathbb{Q}|\mathbb{Q}^{2n}\right) \,:\, \mathbb{Q} \in \mathcal{P}(\mathrm{C}([0,T], \mathcal{M})), \mathbb{Q}_T = p_{\mathrm{ref}}\},$$
$$\mathbb{Q}^{2n+2} = \arg\min\{\mathrm{KL}\left(\mathbb{Q}|\mathbb{Q}^{2n+1}\right) \,:\, \mathbb{Q} \in \mathcal{P}(\mathrm{C}([0,T], \mathcal{M})), \mathbb{Q}_0 = p_0\}.$$

Under mild assumptions on $\mathbb{P}$, p_0 and p_{ref}, we have that $(\mathbb{Q}^n)_{n\in\mathbb{N}}$ converges towards $\mathbb{Q}^\star$ (see Nutz and Wiesel, 2022). In what follows, we propose an algorithm to approximately sample from $(\mathbb{Q}^n)_{n\in\mathbb{N}}$. In Euclidean state spaces, De Bortoli et al. (2021); Vargas et al. (2021); Chen et al. (2021a) have proposed an algorithm based on time-reversal to compute the IPF. We now extend these techniques to the case of Riemannian manifolds.

Riemannian Diffusion Schrödinger Bridge We propose Riemannian Diffusion Schrödinger Bridge (RDSB) an extension of Diffusion Schrödinger Bridge De Bortoli et al. (2021) to approximate solutions of SB problems. First, we connect the iterates $(\mathbb{Q}^n)_{n\in\mathbb{N}}$ with diffusion processes on $\mathcal{M}$.

Proposition S20. *Let $\mathbb{P}$ be the path measure of the Brownian motion initialized at p_{ref}. Assume that for any $n \in \mathbb{N}$, $\mathrm{KL}(\mathbb{Q}^n|\mathbb{P}) < +\infty$ and that for any $t \in [0,T]$ and $n \in \mathbb{N}$, $\mathbb{Q}_t^n$ admits a smooth positive density w.r.t. p_{ref}. Then, for any $n \in \mathbb{N}$ we have:*

(a) $R(\mathbb{Q}^{2n+1})$ solves the martingale problem with generator $\mathcal{A}^{2n+1}(t,u) = \langle \nabla u, b_{T-t}^n \rangle + (1/2)\Delta u$;

(b) $\mathbb{Q}^{2n+2}$ solves the martingale problem with generator $\mathcal{A}^{2n+2}(t,u) = \langle \nabla u, f_t^{n+1} \rangle + (1/2)\Delta u$;

where for any $n \in \mathbb{N}$, $t \in [0,T]$ and $x \in \mathbb{R}^d$, $b_t^n(x) = -f_t^n(x) + \nabla \log p_t^n(x)$, $f_t^{n+1}(x) = -b_t^n(x) + \nabla \log q_t^n(x)$, with $f_t^0(x) = 0$, and p_t^n, q_t^n the densities of $\mathbb{Q}_t^{2n}$ and $\mathbb{Q}_t^{2n+1}$.

Proof. The proof is similar to De Bortoli et al. (2021, Proposition 6) using Theorem 5 instead of Cattiaux et al. (2021, Theorem 4.19) □

In particular, we have that $\mathbb{Q}^1$ is the diffusion process associated with RSGM, *i.e.* the time-reversal of the Brownian motion initialized at p_{ref}. Hence, $\mathbb{Q}^{2n+1}$ for $n \in \mathbb{N}$ with $n \geq 1$ can be seen as a refinement of $\mathbb{Q}^1$. In the next proposition, we show that the drift term of the diffusion processes associated with $(\mathbb{Q}^n)_{n\in\mathbb{N}}$ can be approximated leveraging score-based techniques.

Proposition S21. *Let $(\mathbf{X}_t)_{t\in[0,T]}$ be a $\mathcal{M}$-valued process with distribution $\mathbb{P} \in \mathcal{P}(\mathrm{C}([0,T],\mathcal{M}))$ such that for any $t \in [0,T]$, $\mathbf{X}_t$ admits a positive density $p_t \in \mathrm{C}^\infty(\mathcal{M})$ w.r.t. p_{ref}. Let $s: [0,T] \to \mathcal{X}(\mathcal{M})$. For any $t \in [0,T]$ and $x \in \mathcal{M}$, let*

$$r(t,x) = -s(t,x) + \nabla \log p_t(x).$$

Then, for any $t \in [0,T]$, we have that

$$r(t,\cdot) = \arg\min\{\mathbb{E}[(1/2)\|s(t,\mathbf{X}_t) + r(\mathbf{X}_t)\|^2 + \mathrm{div}(r)(\mathbf{X}_t)] \,:\, r \in \mathrm{L}^2(\mathbb{P}_t)\}.$$

Proof. Let $t \in [0,T]$. First, we have for any $x \in \mathcal{M}$

$$\begin{aligned}&\|r(t,x) - \{-s(t,x) + \nabla \log p_t(x)\}\|^2 \\ &\quad = \|r(t,x) + s(t,x)\|^2 - 2\langle r(t,x), \nabla \log p_t(x)\rangle + \|\nabla \log p_t(x)\|^2 - 2\langle s(t,x), \nabla \log p_t(x)\rangle.\end{aligned}$$

Hence, we get that $r(t,\cdot) = \arg\min\{\mathbb{E}[\|s(t,\mathbf{X}_t) + r(\mathbf{X}_t)\|^2 - 2\langle r(\mathbf{X}_t), \nabla \log p_t(\mathbf{X}_t)\rangle] \,:\, r \in \mathcal{X}(\mathcal{M})\}$. Using the divergence theorem (see Lee, 2018, p.51), we have for any $r \in \mathcal{X}(\mathcal{M})$

$$\begin{aligned}\mathbb{E}[\langle r(\mathbf{X}_t), \nabla \log p_t(\mathbf{X}_t)\rangle] &= \textstyle\int_{\mathcal{M}} \langle r(x_t), \nabla \log p_t(x_t)\rangle p_t(x_t) \mathrm{d}p_{\mathrm{ref}}(x_t) \\ &= -\textstyle\int_{\mathcal{M}} \mathrm{div}(r)(x_t) p_t(x_t) \mathrm{d}p_{\mathrm{ref}}(x_t) = -\mathbb{E}[\mathrm{div}(r)(\mathbf{X}_t)],\end{aligned}$$

which concludes the proof. □

S7 Proof of Proposition 7

Proof. Let $t \in (0,T]$ and $s_t \in \mathrm{C}^\infty(\mathcal{M})$. Using the divergence theorem (see Lee, 2018, p.51), we have

$$\begin{aligned}\ell_{t|s}(s_t) = &\textstyle\int_{\mathcal{M}\times\mathcal{M}} \|\nabla \log p_{t|s}(x_t|x_s)\|^2 \mathrm{d}\mathbb{P}_{s,t}(x_s,x_t) + \int_{\mathcal{M}} \|s_t(x_t)\|^2 \mathrm{d}\mathbb{P}_t(x_t) \\ &\textstyle -2\int_{\mathcal{M}\times\mathcal{M}} \langle \nabla \log p_{t|s}(x_t|x_s), s_t(x_t)\rangle \mathrm{d}\mathbb{P}_{s,t}(x_s,x_t) \\ = &\textstyle\int_{\mathcal{M}\times\mathcal{M}} \|\nabla \log p_{t|s}(x_t|x_s)\|^2 \mathrm{d}\mathbb{P}_{s,t}(x_s,x_t) + \int_{\mathcal{M}} \|s_t(x_t)\|^2 \mathrm{d}\mathbb{P}_t(x_t) \\ &\textstyle -2\int_{\mathcal{M}\times\mathcal{M}} \langle \nabla \log p_{t|s}(x_t|x_s), s_t(x_t)\rangle p_{t|s}(x_t|x_s) p_s(x_s) \mathrm{d}(p_{\mathrm{ref}} \otimes p_{\mathrm{ref}})(x_s,x_t) \\ = &\textstyle\int_{\mathcal{M}\times\mathcal{M}} \|\nabla \log p_{t|s}(x_t|x_s)\|^2 \mathrm{d}\mathbb{P}_{s,t}(x_s,x_t) + \int_{\mathcal{M}} \|s_t(x_t)\|^2 \mathrm{d}\mathbb{P}_t(x_t) \\ &\textstyle -2\int_{\mathcal{M}} \{\int_{\mathcal{M}} \langle \nabla p_{t|s}(x_t|x_s), s_t(x_t)\rangle \mathrm{d}p_{\mathrm{ref}}(x_t)\} p_s(x_s) \mathrm{d}p_{\mathrm{ref}}(x_s) \\ = &\textstyle\int_{\mathcal{M}\times\mathcal{M}} \|\nabla \log p_{t|s}(x_t|x_s)\|^2 \mathrm{d}\mathbb{P}_{s,t}(x_s,x_t) + \int_{\mathcal{M}} \|s_t(x_t)\|^2 \mathrm{d}\mathbb{P}_t(x_t) \\ &\textstyle +2\int_{\mathcal{M}} \{\int_{\mathcal{M}} \mathrm{div}(s_t)(x_t) p_{t|s}(x_t|x_s) \mathrm{d}p_{\mathrm{ref}}(x_t)\} p_s(x_s) \mathrm{d}p_{\mathrm{ref}}(x_s),\end{aligned}$$

which concludes the proof. □

S8 Experimental detail

In what follows we describe the experimental settings used to generate results introduced in Section 6.

Architecture The architecture of the score network s_θ is given by a multilayer perceptron with 5 hidden layers with 512 units each. We use on sinusoidal activation functions.

Optimization All models are trained by the stochastic optimizer Adam (Kingma and Ba, 2015) with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, batch-size of 512 data-points and a learning rate set to $2e-4$.

Likelihood evaluation We rely on the Dormand-Prince solver (Dormand and Prince, 1980), an adaptive Runge-Kutta 4(5) solver, with absolute and relative tolerance of $1e-5$ to compute approximate numerical solutions of the ODE.