RIEMANNIAN NEURAL SDE: LEARNING STOCHASTIC REPRESENTATIONS ON MANIFOLDS

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, the neural stochastic differential equation (NSDE) has gained attention in modeling stochastic representations, while NSDE brings a great success in various types of applications. However, it typically loses the expressivity when the data representation is manifold-valued. To overcome such an issue, we suggest a principled way to express the stochastic representation with the *Riemannian neural SDE* (RNSDE), which extends the conventional Euclidean NSDE. Empirical results on the density estimation on manifolds show that the proposed method significantly outperforms baseline methods.

1 INTRODUCTION

Recently, there has been a great success in modeling stochastic dynamical systems for complex data representations, containing spatially high stochasticity. Especially, the *stochastic differential equation* (SDE) was adopted by recent studies as a fundamental probabilistic model to express the transition of stochastic states. For example, Li et al. (2020) suggested the *neural SDE*, which was the firstly implemented neural network to train the stochastic dynamics and Song et al. (2020) adopted the parameterized reverse-SDE (Anderson (1982)) to model the score-based generative models. Park et al. (2022) introduced the controlled SDE combined with the stochastic optimal control-based theoretical framework to model the time series. These works opened a new way to express the time-series data by showing the power of stochastic representations induced by SDE models. Despite the remarkable recent progress, the major interest in using SDE has been focused on the Euclidean geometry (*i.e.*, $\mathbb{R}^d$). In other words, conventional approaches inevitably lose the expressivity when the data representation us defined on other geometry such as Riemannian manifolds (*i.e.*, $\mathcal{M}$).

In this paper, we tackle the above concern by introducing the *Riemannian neural SDE* (RNSDE) to model the stochastic dynamics on manifolds. The proposed RNSDE is the natural extension of conventional Euclidean SDEs, which defines the intrinsic stochastic transition in the local sense and fully recognizes the geometric structure. Thus, it can resolve expressivity problems of conventional SDEs on a fundamental level. Specifically, our RNSDE is built upon the *Eells-Elworthy-Malliavin* interpretation, where the stochastic trajectories are expressed under the frame bundle geometry.

Contribution. The main contribution of this work is to suggest a general framework for modeling stochastic representations on Riemannian manifolds. Unlike prior works, our objective function enables the proposed model to learn stochastic processes without requiring prior information on target distributions, and thus applicable for high-dimensional and complex datasets. Moreover, by handling Fokker-Planck-Equations, our method can use the network outputs in a direct manner so that its stochastic representation power is strengthened. Experimental results demonstrate that the proposed framework outperforms conventional approaches on the density estimation.

2 RELATED WORK

Normalizing Flows on Riemannian Manifolds. MCNF Lou et al. (2020) and RODE Mathieu & Nickel (2020) defined the ordinary differential equation (ODE) on manifolds and adopted the continuous normalizing flows (*i.e.*, RCNF) to express the transition of data representations. In their methods, the stochastic transition was expressed as the ODE flow of log-density on manifolds and directly calculated the geometric operations (*e.g.*, divergence). In a similar context, diffeomorphic ODE-based flows were adopted to Lie structures (*i.e.*, $SU(n)$) by Katsman et al. (2021), which can preserve the equivariance/invariance.

Riemannian Optimal Transport. Cohen* et al. (2021) directly parameterized the convex potential map on Riemannian manifolds with a neural network, which can solve the optimal transport problem. Rozen et al. (2021) adopted the computationally tractable linear-type normalizing equation to express the evolution of density according to time.

3 RIEMANNIAN NEURAL STOCHASTIC DIFFERENTIAL EQUATION

Notations. Let $\mathcal{M}$ be complete and connected n-dimensional Riemannian manifolds equipped with the Riemannian metric expressed as the matrix form $G \triangleq [g_{ij}]_{1\leq i,j\leq n}$. Similarly, we define the inverse of the Riemannian metric (*i.e.*, co-metric) as $G^{-1} \triangleq [g^{ij}]$. The Christoffel symbol with respect to the metric g is denoted as $[\Gamma^k_{jk}]_{1\leq i,j,k\leq d}$. The orthonormal frame bundle and the tangent bundle are denoted as $O\mathcal{M}$ and $T\mathcal{M}$, respectively. Please refer to Émery (2012); Ikeda & Watanabe (2014); Lee & Lee (2003) for more details on Riemannian geometry and stochastic analysis.

Riemannian Neural SDE. Our major interest is to learn the parameterized stochastic process X^θ_t with parameter θ, which is defined as a solution to the following Stratonovich SDE, Oksendal (1992):

$$dX^\theta_t = \underbrace{W(t, X_t; \theta)dt}_{\text{(A)}} + \underbrace{\left[\int_{s\in\mathbb{T}} \mathcal{T}^t_s W(s, y_s; \theta)ds\right] dt}_{\text{(B)}} + \underbrace{\beta(\theta)\pi^{-1}\left(X^\theta_t\right) \circ dB_t}_{\text{(C)}}, \tag{1}$$

where $B_t = [B^1_t, \cdots, B^n_t]$ denotes the n-dimensional standard Wiener process on $\mathbb{R}^n$. We call the combination of three parameterized terms (*i.e.*, **(A), (B)** and **(C)**) a *Riemannian Neural Stochastic Differential Equation* (RNSDE). The detailed explanations are provided as follows:

(A) Neural Potential Field. In the first term, we call the parameterized vector field $W(t, X_t; \theta) \triangleq w^j(t, X_t; \theta)\partial^t_j : [0, T] \times \mathcal{M} \times \Theta \to T_{X_t}\mathcal{M}$ as the *neural potential field*, where the set of orthonormal tangent vectors $\{\partial^t_j\}_{1\leq j\leq n} \triangleq \{\frac{\partial}{\partial x_j}|_{X_t}\}_{1\leq j\leq n}$ denotes the moving frames defined on the trajectories of stochastic dynamics. This term defines the driving force of stochastic dynamics X^θ_t regarding the inferred network decisions $W(t, X_t; \theta)$ on the current state (t, X_t).

(B) Self-regressive Potential Field. The second term, which is called a *self-regressive potential* (of the kth order), is the main object that characterizes our model. The core idea of this term is to accumulate the decisions of the neural potential field on the history of observed data $\{s, y_s\}_{s\in[t-k,t)}$ from $s = [t - k]$ to $s = t$. Specifically, network decisions $W(s, y_s; \theta) \in T_{X_s}\mathcal{M}$ on past observed data at time $s \in \mathbb{T} \triangleq [t - k, t)$ are translated to the tangent space of the current state by using parallel transport $\mathcal{T}^t_s : T_{X_s}\mathcal{M} \to T_{X_t}\mathcal{M}$, and the translated decisions are accumulated overall. Without loss of generality, we denote this term as $\mathcal{T}(t; \theta)$ in this paper.

(C) Stochastic Development. The last term, called *stochastic development* (Hsu (2002); Ikeda & Watanabe (2014)) controls the diffusive behavior of the proposed SDE on manifolds. The function $\pi : O\mathcal{M} \to \mathcal{M}$ is the canonical projection from frame bundle to model manifolds, and $U_t \triangleq \pi^{-1}(X^\theta_t) \in O\mathcal{M}$ is the *horizontal lift* of our stochastic dynamics $X^\theta_t \in \mathcal{M}$. The parameterized scalar-valued function $\beta : \Theta \to \mathbb{R}^+$ controls the diffusivity of process X^θ_t. It is noteworthy that the solution to (1) is called a Brownian motion on $\mathcal{M}$, when $W \equiv 0$ and $\beta \equiv \frac{1}{2}$.

Local Representations. Although the RNSDE defined in (1) provides the rigorous stochastic representation on manifolds, it is not directly applicable for implementation due to the abstract formulations. Thus, for clarity, we provide the explicit formulation in local representations as follows:

Proposition 1. *(Local Representations) Let X_t be a stochastic process represented as a local coordinate $X_t \triangleq \{x^i_t(w)\}_{1\leq i\leq n} \subset \mathcal{M}$. Then, the proposed RNSDE is defined as follows:*

$$X^\theta_t = \beta(\theta)\int_0^t \sqrt{g^{ij}(X_s)}dB^j_s + \int_0^t \left[\frac{-\beta(\theta)}{2} g^{jk}(X_t)\Gamma^i_{jk}(X_t) + w^j(X_t; \theta)\partial^t_j + \int_{\mathbb{T}} \mathcal{T}^t_s w^j(y_s; \theta)\partial^s_j ds\right] dt, \tag{2}$$

The detailed derivation is provided in Appendix A.1. The intrinsic local coordinate-based representation in (2) only requires the metric tensor g and the Christoffel symbol $\{\Gamma^i_{jk}\}_{1\leq i,j,k,\leq n}$ for every point $x \in \mathcal{M}$, where the newly appearing terms (*i.e.*, first and second) in (2) reveal the effect of stochastic development. It transforms the Euclidean stochasticity induced by standard Brownian motion B_t into Riemannian stochasticity on manifolds by considering the curvature effect (*i.e.*, g, Γ).

As the propagation of stochastic particles $\{X_t\}_{0\leq t\leq T}$ randomly changes the moving frame $\{\partial_j(X_t)\}_{1\leq j\leq d}$ and its local coordinate system, we call the solution to (2), X_t, as the *stochastic local chart flow* (SLCF). In this paper, we regard that the Riemannian manifolds M are embedded

compact sub-manifolds of the ambient space, $\mathcal{M} \subseteq \mathbb{R}^D$ for $n \leq D$. In other terms, the coordinate system of the diffusion process X_t can be expressed in $\mathbb{R}^D$ using the global coordinate as $Y_t \triangleq \psi \circ X_t$ where ψ is the coordinate function.

4 Learning Densities from Samples

Learning Densities from Samples. The major drawback of prior works is that full information on target probability densities is required. For example, the conventional RCNF optimizes log-likelihood (or KL-divergence) between source and target distributions, *i.e.*, $\min_\theta \mathbb{E}[\log p_\theta - \log p_\nu]$. Unfortunately, this formulation may cause problems in high-dimensional and complex datasets, as one needs to priorly access the accurate estimation of the data density p_ν (*e.g.*, KDE). In contrast, our method utilizes only the set of observed particles $\{y_t\} \sim \nu_t$ sampled from the unknown probability measure ν_t. In other words, we assume that no prior information on target distributions is given.

Objective Function. From the N-number of observed particles $\{y_t^l\}_{1 \leq l \leq N}$ at time t, we define the empirical target measure as $\nu_t = \frac{1}{N}\sum_l^N \delta_{y_t^l}$ to approximate. Then, the source probability measure is defined as the law of solution to the proposed RNSDE: $\mu_t^\theta \triangleq \mathbb{P}(X_t^\theta \in \cdot) \sim X_t^\theta \in \mathcal{M}$. Note that the source measure $\mu_t^\theta | \{\mathbf{y}_s\}_{t-k \leq s < t}$ is conditioned by the set of conditional observations $\{\mathbf{y}_s\}_{t-k \leq s < t}$ encoded in the self-regressive potential field, *i.e.*, (1)-**(B)**. Regarding the definition of μ_t^θ and ν, our objective function is posed to minimize the distributional discrepancy (*i.e.*, Wasserstein distance) between source and target measures:

$$\min_\theta \mathcal{W}_2\left(\mu_t^\theta | \{\mathbf{y}_s\}_{t-k \leq s < t},\ \nu_t\right). \tag{3}$$

Specifically, our objective function is posed to minimize the 2-Wasserstein distance at time t. Then, the problem is accurately calculating the Wasserstein distance by relating the probabilistic structure of RNSDE. To do so, we are especially interested in the Markovian property:

Markov Diffusive Kantorovich Dual. By the independent increments of B_t, one can show that the proposed RNSDE preserves the Markov property. Let $P_t^\theta f \triangleq \mathbb{E}[f(X_t^\theta)|X_0]$ be a Markov semi-group of X_t^θ. Then, by duality, one can define the probability measure called *dual semi-group* $P_t^{\theta,*}$ satisfying the equality, $\mathbb{E}_{x\sim\mu_0}[P_t^\theta f(x)] = \mathbb{E}_{x'\sim P_t^{\theta,*}\mu_0}[f(x')]$, which identifies source probability measure as $\mu_t^\theta \triangleq P_t^{\theta,*}$. This theoretical interpretation has been proposed by Park et al. (2021) to express the temporal path of probability measure for the Markov process. Inspired by their work, we define the dual Kantorovich formulation with the dual semi-group to define the Wasserstein distance:

Definition 1. *(Markov Diffusive Kantorovich Dual) Let us define two arbitrary functions $A, B \in C^2(\mathcal{M})/\mathbb{R}$. Then, the diffusive Kantorovich dual is defined as follows:*

$$\mathcal{J}^\epsilon([A,B],\theta) = \int A(x)dP_t^{\theta,*}(x) + \int B(y)d\nu_t(y) - \epsilon \int e^{\frac{A(x)+B(y)-d^2(x,y)}{\epsilon}} dP_t^{\theta,*} \otimes \nu_t. \tag{4}$$

It is well-known that the fixed point $[A^\star, B^\star]$ of *Sinkhorn-iteration* started at $[A, B]$ uniquely determines the equality, $\lim_{\epsilon\to 0}\mathcal{J}^\epsilon([A^\star,B^\star];\theta) = \mathcal{W}_2(\mu_t^\theta, \nu_t)$. Using this relation, our objective function in (3) can be reformulated as $\min_\theta \mathcal{J}^{\Delta\epsilon}([A^\star,B^\star];\theta)$ with pre-determined parameter $\epsilon \triangleq \Delta\epsilon \approx 0$. To solve this new minimization problem, we apply the gradient descent as follows:

$$\theta_{m+1} = \theta_m - \tau_m \mathbb{E}\Bigg[\partial_\theta \beta(\theta) \int_0^t g^{jk}\partial_{j,k}A^\star(X_s^\theta) - \partial_\theta w^j(\cdot;\theta)\partial_j A^\star(X_s^\theta) - \partial_\theta \sum_{s\in\mathbb{T}} \mathcal{T}_s^t w^j(y_s;\theta)\partial_j^s A^\star(X_s^\theta)ds\Bigg], \tag{5}$$

where τ_m is the learning rate. Please note that the expectation is taken to measure $\mathbb{P}$. Detailed derivations are provided in Appendix A.3.

4.1 Comparison to Existing Methods

In conventional Riemannian normalizing flows, the model density p_θ is transited by following the propagation rule: $\partial_t \log p_\theta(z_t) = -\mathbf{div}_{\mathcal{M}}[V_\theta(z_t,t)]$, where z_t denotes the ODE flow on the manifold, and V_θ is the parameterized vector field by the neural network. For the numerical simulation of the

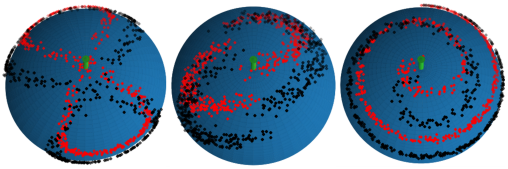

Figure 1: **MCNF**, Lou et al. (2020)

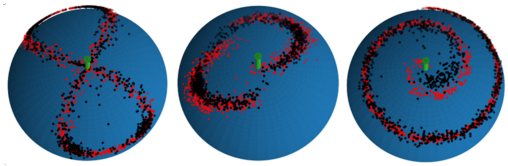

Figure 2: **RNSDE** (ours)

equation, they require the computation of Hessian trace. In our case, the model density p_t is defined as the solution to the second-order PDE called a *parameterized Fokker-Planck Equation, (pFPE)*:

$$\textbf{pFPE:} \quad \partial_t p(t,x) = \beta(\theta)\Delta_{\mathcal{M}} p(t,x) - \mathbf{div}_{\mathcal{M}}[p(t,x)V_\theta(t,x)], \tag{6}$$

where $V_\theta(t,\cdot) \triangleq W(t,\cdot;\theta) + \mathcal{T}(t;\theta)$. In contrast to the conventional approach, the proposed RNSDE does not need to simulate the pFPE directly. As shown in (4), we indirectly access the measure μ_t^θ by calculating the Markovian semi-group. Thus, our method can avoid the computational burden to calculate the geometric operations for normalizing flows (*e.g.*, divergence, log-determinant), while preserving the representational power. Katsman et al. (2021) induce the vector field $V(t,\cdot) \triangleq \nabla \log(p_\nu(t,\cdot)/p_\theta(t,\cdot)) \in T\mathcal{M}$, which also solves the pFPE. However, the representational power can be limited as the network outputs (*i.e.*, Neural ODE) are implicitly utilized to calculate model density p_θ. In contrast, our method directly utilizes the network outputs (*e.g.*, $W, \mathcal{T} \in T\mathcal{M}$) to model the vector field V in (6), which enables rich stochastic representations.

5 Experiments

In the experiment, we evaluated the performance for estimating densities on the 2-sphere, including three different densities: *8-shapes, Two moons, and Spiral*. Note that the complexity of these densities is higher compared to those of previous works (*e.g.*, a mixture of von Mises), and the density estimation is more challenging.

To define the target densities for baseline models, including MCNF (Lou et al. (2020)) and EMSRE (Rezende et al. (2020)), we utilized samples from target densities as the anchor points of Gaussian radial functions. After defining the approximated target densities obtained from the Gaussian KDE, both models were trained to minimize the KL divergence. For our model, we initialized the start point of stochastic trajectories as $X_0^\theta \sim \boldsymbol{\mu}_0 \triangleq \psi_\# \mathcal{N}(0, \mathbf{I}_n)$, where $\mathcal{N}$ denotes the Gaussian distribution and ψ is coordinate function. For self-regressive behavior, the conditional observations are defined as $y_s \sim \psi_\# \mathcal{N}(\mathbb{E}[X_s^\theta], \mathbf{I}_n)$. As a consequence, the information of past trajectories $X_{\{s<t\}}^\theta$ is consecutively encoded in the current state. For the RNODE, which is a deterministic version of our RNSDE, we utilized only the potential field term in (1)-(A), where other terms were set to 0.

Table 1: **Performance Comparison of Density Estimation.** Each model was evaluated in terms of the 2-Wasserstein distance $\mathcal{W}_2$ ($\times 10^{-2}$).

Methods	MCNF	EMSRE ($N_T = 24, K = 5$)	RNODE	RNSDE
8-shapes	11.258	9.826	8.007	**6.052**
Two moons	14.335	9.110	7.684	**5.871**
Spiral	15.153	10.129	9.316	**7.294**

Experimental Results. As shown in Table 1, the proposed RNSDE significantly outperforms other baselines by a large margin. This highlights the representation power of the proposed RNSDEs. The RNODE shows an inferior outcome compared to the RNSDE, which shows the effectiveness of providing additional information (*i.e.*, conditional observations in (1)-(B) and (C)). Figs 1 and 2 display samples from learned densities (*i.e.*, black dots) and target densities (*i.e.*, red dots) for the baseline and our model, respectively. While the MCNF fails to approximate the target densities due to the complex geometric shapes, our method can fit the target densities accurately.

6 Conclusion

In this paper, we suggested a principled way to express the stochastic representations on manifolds. Specifically, we introduced the Riemannian neural SDE with stochastic development defined on an orthonormal frame bundle. In future work, we plan to theoretically extend the current work by analyzing the stability of Sinkhorn-iteration and show the learnability of the RNSDE.

REFERENCES

Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Robert J Berman. The sinkhorn algorithm, parabolic optimal transport and geometric monge–ampère equations. *Numerische Mathematik*, 145(4):771–836, 2020.

Samuel Cohen*, Brandon Amos*, and Yaron Lipman. Riemannian Convex Potential Maps. In *ICML*, 2021.

Michel Émery. *Stochastic calculus in manifolds*. Springer Science & Business Media, 2012.

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

Nobuyuki Ikeda and Shinzo Watanabe. *Stochastic differential equations and diffusion processes*. Elsevier, 2014.

Isay Katsman, Aaron Lou, Derek Lim, Qingxuan Jiang, Ser-Nam Lim, and Christopher De Sa. Equivariant manifold flows. In *NeurIPS*, 2021.

J.M. Lee and J.M. Lee. *Introduction to Smooth Manifolds*. Graduate Texts in Mathematics. Springer, 2003. ISBN 9780387954486.

Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David Duvenaud. Scalable gradients for stochastic differential equations. In *AISTATS*, 2020.

Aaron Lou, Derek Lim, Isay Katsman, Leo Huang, Qingxuan Jiang, Ser-Nam Lim, and Christopher De Sa. Neural manifold ordinary differential equations. In *NeurIPS*, 2020.

Emile Mathieu and Maximilian Nickel. Riemannian continuous normalizing flows. *NeurIPS*, 2020.

Robert McCann. Polar factorization of maps on riemannian manifolds. *Geometric and Functional Analysis*, 11:589–608, 08 2001. doi: 10.1007/PL00001679.

Bernt Oksendal. *Stochastic Differential Equations : An Introduction with Applications*. Springer-Verlag, Berlin, Heidelberg, 1992.

Sung Woo Park, Dong Wook Shu, and Junseok Kwon. Generative adversarial networks for markovian temporal dynamics: Stochastic continuous data generation. In *ICML*, 2021.

Sung Woo Park, Kyungjae Lee, and Junseok Kwon. Neural markov controlled SDE: Stochastic optimization for continuous-time data. In *ICLR*, 2022.

Danilo Jimenez Rezende, George Papamakarios, Sebastien Racaniere, Michael Albergo, Gurtej Kanwar, Phiala Shanahan, and Kyle Cranmer. Normalizing flows on tori and spheres. In *ICML*, 2020.

Noam Rozen, Aditya Grover, Maximilian Nickel, and Yaron Lipman. Moser flow: Divergence-based generative modeling on manifolds. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2020.

François-Xavier Vialard. An elementary introduction to entropic regularization and proximal methods for numerical optimal transport. 2019.

C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008.

Feng-Yu Wang. *Analysis for diffusion processes on Riemannian manifolds*, volume 18. World Scientific, 2014.

A APPENDIX

A.1 BASICS OF STOCHASTIC DIFFERENTIAL GEOMETRY

In the orthonormal frame bundle $O\mathcal{M}$, the local coordinate is defined as $U_t = (X_t, E_j^k) \in O\mathcal{M}$, where the tangent of frame bundle can be decomposed as direct sum of two subspaces: $TO\mathcal{M} = VO\mathcal{M} \oplus HO\mathcal{M}$.

$$\begin{aligned} dX_t^{i,\theta} &= \sum_j^d E_j^i(U_t) \circ dB_t^j + V_\theta(X_t)dt, \\ dE_j^i(t) &= -\Gamma_{kl}^i E_j^l(t) \circ dX_t^k, \end{aligned} \tag{7}$$

where $V_\theta(t,\cdot) \triangleq W(t,\cdot;\theta) + \mathcal{T}(t;\theta)$. In full description, the Stratonovich SDE in the orthonormal frame bundle is written as the semi-martingale:

$$dU_t = \underbrace{H_i(U_t) \circ dB_t^i}_{\text{Local martingale}} + \underbrace{\hat{V}_\theta(U_t)dt}_{\text{Adapted process}}, \tag{8}$$

where $\hat{V}_\theta$ is a horizontal lift of vector field V_θ. By applying Itó's lemma to the SDE in (8) with the function $f = R \circ \pi$ for arbitrary $R : \mathcal{M} \to \mathbb{R}$, one can obtain the following representation of SDE.

$$f(U_t) = f(U_0) + \int_0^T H_i f(U_t) \circ dB_t^i + \int_0^T \hat{W}_\theta f(U_t)dt. \tag{9}$$

In the local coordinate presentation, we can written (9) as follows:

$$\begin{aligned} dR(X_t) &= H_i R(X_t) \circ dB_t^i + W(X_t;\theta)R(X_t)dt \\ &= H_i R(X_t)dB_t^i + \frac{1}{2}d[H_i R, B]_t + \mathbb{W}_t^\theta \\ &= R(X_0) + \mathbb{M}_t + \frac{1}{2}\mathbb{N}_t + \mathbb{W}_t^\theta. \end{aligned} \tag{10}$$

As the lifted function $f = R \circ \pi$ uniquely determines the orthonormal basis of tangent space $ue_i \in T_\mathcal{M}$ by the property of the fundamental horizontal vector field H_i (*i.e.*, there exists a unique relation $\pi_\star H_i(U_t) = U_t e_i$), we can express $H_i f(U_t) = U_t e_i$. This fact leads the second equality in the following formulation:

$$\mathbb{M}_t = \sum_i H_i f(U_t) dW_t^i = \sum_i (df)_{U_t}[H_i(U_t)]dW_t^i. \tag{11}$$

To understand the above relation precisely, $\mathbb{M}_t$ can be expressed as local coordinate. For this, we denote the basis of horizontal vector space $HO\mathcal{M}$ as $\{\partial_j \triangleq \frac{\partial}{\partial x_j}, \bar{\partial}_m^k \triangleq \frac{\partial}{\partial e_m^k}; 1 \le j,k,m \le d\}$. In this case, the horizontal curve U_t in orthonormal frame bundle $O\mathcal{M}$ can be represented in local coordinates, $U_t = [X_t^i, e_j^i(t)]$ as follows:

$$\begin{aligned} H_i(U_t)[f] &\triangleq H_i f(U_t) = (df)_{U_t}[H_i(U_t)] \\ &= e_i^j \partial_j f(U_t) - e_i^j e_m^l \Gamma_{jl}^k(X_t)\bar{\partial}_m^k f(U_t) \\ &= e_i^j \partial_j R \circ \pi([X_t^i, e_j^i(t)]) \\ &= (e_i^j \partial_j)_{X_t} R(X_t^i) = (U_t e_i)_{X_t} R(X_t^i) \\ &= T(X_t^i) R(X_t^i), \end{aligned} \tag{12}$$

where $H_i(U_t) \triangleq e_i^j \partial_j - e_i^j e_m^l \Gamma_{jl}^k(X_t)\bar{\partial}_m^k$ is the fundamental horizontal vector field. The third equality holds as $\bar{\partial}_m^k \circ f([X_t^i, e_j^i(t)]) = \bar{\partial}_m^k R \circ \pi([X_t^i, e_j^i(t)]) = \bar{\partial}_m^k R(X_t) = 0$. In the last equality, we define the vector field $T \in T\mathcal{M}$ as $ue_i = e_i^j \partial_j \triangleq T(X_t)$. To estimate the vector field $T(X_t)$, we need to find out the explicit numeric of the orthogonal matrix e_i^j by solving the following equation:

$$(Ue_i)|_{X_t} = e_i^j \partial_j|_{X_t}. \tag{13}$$

To solve the equation, we take the Riemannian inner product between Ue_l and Ue_m, as follows:

$$\begin{aligned}\langle Ue_l, Ue_m\rangle_{g(X_t)} &= \langle e_l^i\partial_i, e_m^j\partial_j\rangle_{g(X_t)} = e_l^i\langle\partial_i,\partial_j\rangle_{g(X_t)}e_m^j \\ &= e_l^i g_{ij}(X_t)e_m^j = \delta_{ij},\end{aligned} \tag{14}$$

where we denote $\langle X, Y\rangle_g$ as the inner product between vector fields $X, Y \in T\mathcal{M}$ and δ_{lm} is the coordinate delta. Using the relation in (14), the following identity can be easily obtained:

$$\sum_k e_k^i e_k^j = g^{ij}. \tag{15}$$

One may express the identity in (15) as the matrix form, as follows:

$$E^T E = G^{-1}, \tag{16}$$

where we denote $E = \{e_k^i\}$, $G^{-1} = g^{ij}$. To obtain the explicit form of matrix E, we apply the Cholesky decomposition to the co-metric matrix. (*i.e.*, $E = \mathbf{Ch} \circ [G^{-1}]$). Finally, the derivation form of horizontal vector field to f, $H_i f(U_t)$ can be written in local coordinate as follows:

$$H_i f(U_t) = (U_t e_i)|_{X_t} R(X_t^i) = \mathbf{Ch} \circ [G^{-1}(X_t)]^i \partial_i R(X_t^i). \tag{17}$$

As the metric matrix is semi-definite positive in our case, the following equality holds by the elementary algebraic property:

$$\mathbf{Ch} \circ G^{-1} = G^{-\frac{1}{2}}. \tag{18}$$

Finally, the local martingale term $\mathbb{M}_t$ is written in local coordinate as follows:

$$\mathbb{M}_t = \sum_i G(X_t)^{-\frac{1}{2}} \partial_i R(X_t) dW_t^i. \tag{19}$$

Let us calculate the quadratic variation of the process $\mathbb{M}_t$.

$$d\left[\mathbb{M}, \mathbb{M}\right]_t = \left[\sum_i \sqrt{g_{ii}^{-1}}\partial_i R(X_t)\right]^2 dt, \tag{20}$$

where $[B, B]_t = dt$. Thus, $\mathbb{M}_t$ is the local martingale.

$$\begin{aligned}\mathbb{N}_t = [H_i R, B]_t &= \int H_j H_i f(U_t) d\left[B^j, B^i\right]_t \\ &= \int \sum_i H_i^2 f(U_t) dt = \int \Delta_{O\mathcal{M}} f(U_t) = \int \Delta_{\mathcal{M}} R(X_t),\end{aligned} \tag{21}$$

where $\Delta_{O\mathcal{M}} = \sum_i H_i^2$ denotes the Bochner's horizontal Laplacian operator. The last equality is known as the Bochner's Laplacian identity. The last term $\mathbb{W}_t^\theta = V(X_t;\theta)R(X_t)$ corresponds to the anti-development of vector field $\hat{V}_\theta$. By collecting the defined stochastic representations, $\mathbb{M}, \mathbb{N}$, and $\mathbb{W}$, the local expressions in (2) is derived.

A.2 Local coordinate system

Among the various definitions of the local chart ψ, we can select the following two types of local coordinate system:

1) Parameterization. Let (ϑ, φ) be a local coordinate of the 2-sphere. Then, the global expression is parameterized with spherical coordinate in the following form:

$$\psi(\vartheta, \varphi) = [\cos\varphi\sin\vartheta, \sin\varphi\sin\vartheta, \cos\vartheta] \in \mathbb{R}^3. \tag{22}$$

2) Normal Coordinate. While this coordinate system induces the vanishing metric tensor and the Christoffel symbol at X_t (*i.e.*, $g|_{X_t}, \Gamma|_{X_t} = 0$), the diffusion process X_t in the ambient coordinate system on $\mathbb{R}^k$ is expressed as the following form:

$$d_I Y_t = d_I(\psi \circ X_t) = \exp_{X_t}\left(\partial_j dB_t^j + w^j(X_t;\theta)\partial_j dt + \int_{\mathbb{T}} \mathcal{T}_s^t w^j(y_s;\theta)\partial_j^s ds\right). \tag{23}$$

In the literature, the deterministic version of the proposed differential equation in (23) is known as the *exponential-map flows*, Rezende et al. (2020).

A.3 MARKOV DIFFUSIVE KANTOROVICH DUAL FORMULATION

In this section, we introduce the intuition behind the gradient descent suggested in (5). The propositions introduced in this section can be found in Berman (2020) and Vialard (2019).

For each function $A, B \in C(\mathcal{M})^2/\mathbb{R}$, we define two functional operators on $C(\mathcal{M})^2/\mathbb{R}$, as follows:

$$\mathcal{H}^{\epsilon}_{\mu_t} : C(\mathcal{M}) \to C(\mathcal{M}), \quad \mathcal{H}^{\epsilon}_{\mu_t}[A](y) = \epsilon \log \int e^{-d^2(x,y)/\epsilon - A(x)/\epsilon} \mu_t(dx), \tag{24}$$

$$\mathcal{H}^{\epsilon}_{\nu} : C(\mathcal{M}) \to C(\mathcal{M}), \quad \mathcal{H}^{\epsilon}_{\nu}[B](x) = \epsilon \log \int e^{-d^2(x,y)/\epsilon - B(y)/\epsilon} \nu_t(dy). \tag{25}$$

Then, the composition of two operators (*i.e.*, $\mathcal{H}^{\epsilon}_{\mu_t}, \mathcal{H}^{\epsilon}_{\nu}$) is called a *log-Sinkhorn* iteration $\mathcal{S} \triangleq \mathcal{H}^{\epsilon}_{\mu_t} \circ \mathcal{H}^{\epsilon}_{\nu}$. Let us define A_l is the transformed shape of A after l-th iteration: $A_l = \mathcal{S}^l(A) = \underbrace{\mathcal{S} \circ \mathcal{S} \cdots \circ \mathcal{S}}_{l \text{ time}}(A)$. Let us define the functional $\overline{\mathcal{F}}(A_l) : C(\mathcal{M}, \mathbb{R}) \to \mathbb{R}$, as follows:

$$\overline{\mathcal{F}}(A_l) = \int \mathcal{S}^l \circ A(x) d\mu^{\theta}_t(x) + \int \mathcal{H}^{\epsilon}_{\mu_t}[\mathcal{S}^l \circ A](y) d\nu_t(y). \tag{26}$$

Then, the log-Sinkhorn iteration uniquely minimizes the functional $\overline{\mathcal{F}}$ by the following proposition.

Proposition 2. *The log-Sinkhorn iteration $\mathcal{S}$ has a fixed point $A_{l\to\infty}$ in $C(\mathcal{M}, \mathbb{R})/\mathbb{R}$. This fixed point is determined up to additive constant, and minimize the functional $\overline{\mathcal{F}}$ uniformly:*

$$\overline{\mathcal{F}}(\mathcal{S} \circ A_l) \triangleq \overline{\mathcal{F}}(A_{l+1}) \leq \overline{\mathcal{F}}(A_l). \tag{27}$$

Let us assume that, for the large enough $l > L$ with small enough $\epsilon \approx 0$, the log-Sinkhorn iteration is converged $\mathcal{S} \circ A_l = A_l$, and the functional $\overline{\mathcal{F}}$ is minimized. Then, the function $A_{l \vee L}$ is approximated to the d^2-transformation (Villani (2008)) of the function $B_{m \vee M}$.

$$[A_{l \vee L}]^c \approx \left[\chi_{\textbf{supp}(\mu_t)} + \mathbf{B}_{l \vee L}\right]^c, \tag{28}$$

where $\chi_V(x) = 0$, if $x \in V$, and $\chi_V(x) = \infty$ if $x \notin V$ for the compact subset $V \subset \mathcal{M}$, and $[f]^c$ denotes the d^2-transformation of f. By the Kantorovich duality theorem, this theoretical characteristic of log-Sinkhorn iteration induces the optimal transport between μ_t and ν_t.

Theorem 1. *(Unique fixed-point of $\mathcal{S}$ and Optimal transport) Assume that $(\mu^{(\epsilon)}_t, \nu^{(\epsilon)}_t) \to (\mu_t, \nu_t)$ in $\mathcal{P}(\mathcal{M})$-weak sense. If A is the fixed point of the log-Sinkhorn operator $\mathcal{S}$ on $C(\mathcal{M}, \mathbb{R})$, then B converges uniformly to a d^2-convex function such that there exists a unique and optimal transport map Φ satisfying:*

$$\Phi(y) = \exp_y(\nabla_g B(y)), \quad \Phi_{\#}\nu = \mu_t, \tag{29}$$

where $\exp$ and ∇_g are Riemannian exponential and gradient, respectively.

In short, the proposed iteration approximates the d^2-convex function $B \approx A^c$ and solves the optimal transport problem, *i.e.*, 2-Wasserstein distance (McCann (2001)).

Parameterized Generator. Our interest is to investigate the stochastic and geometric effect on evaluating the functional $\overline{\mathcal{F}}$ regarding the probabilistic property of X^{θ}_t. To do so, we first define the second-order partial differential operator called *infinitesimal generator* as follows:

$$\mathcal{L}^{\theta} f(x) \triangleq \beta(\theta) \Delta_{\mathcal{M}} f(x) - (W_{\theta} + \mathcal{T}_{\theta}) f(x), \quad x \in \mathcal{M}, f \in C^2(\mathcal{M}), \tag{30}$$

where $\Delta_{\mathcal{M}} \triangleq \mathbf{div}_{\mathcal{M}} \circ \nabla_{\mathcal{M}}$ is the Laplace-Beltrami operator on $\mathcal{M}$. Note that, due to the time-dependent behavior, the proposed generator induces the time-inhomogeneous Markov process.

Lemma 1. *The solution of the proposed RNSDE, X^{θ}_t, is the Markov process for any $\theta \in \Theta$.*

As the proposed process X^{θ}_t is the Markov diffusion process regarding the definition of generator in (30), one can apply the geometric version of Itó's lemma (Wang (2014)) to obtain the following equality:

$$\mathbf{A}^t_m(X^{\theta}_t; \theta) = A_m(X_0) + \int_0^t \mathcal{L}^{\theta} A_m(X^{\theta}_s) ds + \frac{\beta(\theta)}{2} \int_0^t \langle U_s^{-1} \nabla_g A_m(X_s), dB_s \rangle, \tag{31}$$

where $U_s^{-1} : T_{X_s^\theta}\mathcal{M} \to \mathbb{R}^d$ is the inverse of frame U_s at X_s. Note that the notation for A_m is rewritten in LHS of (31), $\mathbf{A}^\theta \in C(\mathcal{M}, \mathbb{R}) \times \Theta$ to emphasize the parameterized generator $\mathcal{L}^\theta$ in RHS.

$$\begin{aligned}
\overline{\mathcal{F}}(\mathbf{A}_m^t(\cdot,\theta)) &= \underbrace{\int \mathbf{A}_m^t(X_t^\theta,\theta)d\mu_t}_{\text{Parameter Activated}} + \underbrace{\int \mathcal{H}^\epsilon_{\mu_t^{\theta_0}}[A_m](y)d\nu(y)}_{\text{Parameter Frozen}} \\
&= \mathbb{E}_{x_0\sim p_0}\mathbb{E}[\mathbf{A}_m^t(X_t^\theta,\theta)|X_0 = x] + \int \mathcal{H}^\epsilon_{\mu_t^{\theta_0}}[A_m](y)d\nu(y) \\
&= \mathbb{E}_{x_0\sim p_0}\mathbb{E}\left[A_m(x) + \int_0^t \mathcal{L}^\theta A_m(X_s^\theta)ds|X_0 = x\right] + \int \mathcal{H}^\epsilon_{\mu_t^{\theta_0}}[A_m](y)d\nu(y) \\
&= \mathbb{E}_{x_0\sim p_0}\mathbb{E}\left[A_m(x) + \beta(\theta)\int_0^t \Delta_{\mathcal{M}}A_m - (W_\theta + \mathcal{T}_\theta)A_m ds|X_0 = x\right] + \int \mathcal{H}^\epsilon_{\mu_t^{\theta_0}}[A_m](y)d\nu(y).
\end{aligned} \tag{32}$$

Notice. In the first equality, the parameter in the second term $\int \mathcal{H}^\epsilon_{\mu_t^{\theta_0}}[A_m](y)d\nu(y)$ is considered as fixed $\theta = \theta_0$ during the update. This trick makes us to avoid calculating $\partial_\theta \int \mathcal{H}^\epsilon_{\mu_t^{\theta_0}}[A_m](y)d\nu(y)$, which leads unstable results.

Gradient flow (descent) of functional $\overline{\mathcal{F}}$ with respect to Parameter θ. In dual formulation, the gradient of functional $\overline{\mathcal{F}}$ can be written as follows:

$$\partial_\theta\overline{\mathcal{F}}(\mathbf{A}_m^t(\cdot,\theta)) = \partial_\theta\mathbb{E}\mathcal{L}^\theta A = \mathbb{E}\partial_\theta\beta(\theta)A - \mathbb{E}\partial_\theta W_\theta A - \mathbb{E}\partial_\theta\mathcal{T}_\theta A. \tag{33}$$

In local coordinate, the gradient with respect to parameter θ is written as follows:

$$\begin{aligned}
&\partial_\theta\overline{\mathcal{F}}(\mathbf{A}_t(X_t^\theta;\theta)) \\
&= \mathbb{E}\left[\partial_\theta\beta(\theta)\int_0^t g^{jk}\partial_{j,k}A(X_s^\theta)ds - \partial_\theta w^j(\cdot;\theta)\partial_j A(X_s^\theta)ds - \partial_\theta\sum_{s\in\mathbb{T}}\mathcal{T}_s^t w^j(y_s;\theta)\partial_j^s A(X_s^\theta)ds\right],
\end{aligned} \tag{34}$$

where A is the fixed point of log-Sinkhorn iteration. Finally, the gradient flow of θ to minimize the functional $\overline{\mathcal{F}}$ is defined in the following form:

$$d\theta(s)_{0\le s<\infty} = -\tau\partial_\theta\overline{\mathcal{F}}(\mathbf{A}_t(X_t^\theta;\theta(s))). \tag{35}$$

The proposed gradient descent in (5) is temporal discretized version of (35).

A.4 Geometric Calculations

In this section, we provide the detailed geometric calculation on the 2-Sphere, $\mathbb{S}^2$.

Riemannian Metric Tensor, Christoffel Symbol. The following calculation is conducted upon the spherical coordinate system defined in (22). Then, the Riemannian metric and Christoffel symbol is calculated as follows:

$$g_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & \sin^2\vartheta \end{pmatrix}, \quad \Gamma_{jk}^i = \left[\begin{pmatrix} 0 & 0 \\ 0 & -\sin\vartheta\cos\vartheta \end{pmatrix}, \begin{pmatrix} 0 & \cot\vartheta \\ \cot\vartheta & 0 \end{pmatrix}\right], \tag{36}$$

where the Christoffel symbol is defined as $\Gamma_{jk}^i = \frac{1}{2}g^{ke}\left(\partial_j g_{ei} + \partial_i g_{ej} - \partial_e g_{ij}\right)$ and ∂_i denotes the (spatial) partial derivative with respect to the i-th component.

Parallel Transport, Exponential Map. Let U be a tangent vector on $T_{X_s}\mathcal{M}$ at time s, and let X_t be a another point at time t. Then, the parallel transport of tangent vector U from $T_{X_s}\mathcal{M}$ to $T_{X_t}\mathcal{M}$ is defined as follows:

$$\mathcal{T}_s^t(U) = U - \mathbf{Tr}[(X_t)^T U]X_t. \tag{37}$$

Note that the stochastic process X_s is recognize as global representations in the evaluation. For the Riemannian exponential map, we define the following normalized version:

$$\mathbf{Exp}_{X_t}(U) = \cos(\|U\|)X_t + \frac{\sin\|U\|}{\|U\|}U. \tag{38}$$