Convergence of Kinetic Langevin Monte Carlo on Lie groups

Lingkai Kong   LKONG75@GATECH.EDU

Molei Tao   MTAO@GATECH.EDU

Georgia Institute of Technology

Editors: Shipra Agrawal and Aaron Roth

Abstract

Explicit, momentum-based dynamics for optimizing functions defined on Lie groups was recently constructed, based on techniques such as variational optimization and left trivialization. We appropriately add tractable noise to the optimization dynamics to turn it into a sampling dynamics, leveraging the advantageous feature that the trivialized momentum variable is Euclidean despite that the potential function lives on a manifold. We then propose a Lie-group MCMC sampler, by delicately discretizing the resulting kinetic-Langevin-type sampling dynamics. The Lie group structure is exactly preserved by this discretization. Exponential convergence with explicit convergence rate for both the continuous dynamics and the discrete sampler are then proved under W_2 distance. Only compactness of the Lie group and geodesically L-smoothness of the potential function are needed. To the best of our knowledge, this is the first convergence result for kinetic Langevin on curved spaces, and also the first quantitative result that requires no convexity or, at least not explicitly, any common relaxation such as isoperimetry.

Keywords: left-trivialized kinetic Langevin dynamics; momentum Lie group sampler; nonasymptotic error bound; nonconvex; no explicit isoperimetry

1. Introduction

Sampling is a classical field that is nevertheless still rapidly progressing, with new results on quantitative and nonasymptotic guarantees, new algorithms, and appealing machine learning applications. One type of methods is constructed by, or at least interpretable as discretizing a continuous time SDE whose equilibrium is the target distribution. Some famous examples are Langevin Monte Carlo (LMC), which originates from overdamped Langevin SDE, and kinetic Langevin Monte Carlo (KLMC), which originates from kinetic Langevin SDE. Both LMC and KLMC are widely used gradient-based methods. Similar to the fact that momentum helps gradient descent (Nesterov, 2013), KLMC can be interpreted as a momentum version of LMC.

Nonasymptotic error analysis of momentumless Langevin algorithm in Euclidean spaces have be established by a collection of great works (e.g., Dalalyan, 2017; Cheng and Bartlett, 2018; Durmus et al., 2019; Vempala and Wibisono, 2019; Li et al., 2021). Samplers based on kinetic Langevin are more difficult to analyze, largely due to the degeneracy of noise as it is only added to the momentum but not the position, but many remarkable progress (e.g., Shen and Lee, 2019; Dalalyan, 2017; Ma et al., 2021; Zhang et al., 2023; Yuan et al., 2023; Altschuler and Chewi, 2023) have been made too, showcasing the benefits of momentum. More discussion of sampling in Euclidean spaces can be found in Sec. B.1.

Sampling on a manifold that does not admit a global coordinate is much harder. Even the exact numerical implementation of Brownian motion on it is difficult for general manifolds (Hsu, 2002). Seminal results for general Riemannian manifolds with nonasymptotic error guarantees include Cheng et al. (2022); Wang et al. (2020); Gatmiry and Vempala (2022), leveraging either a discretization of the Brownian motion or assuming an oracle that exactly implements Brownian motion (which is feasible for certain manifolds). However, all these results focused on the momentumless case.

For a problem closely related to sampling, namely optimization though, several great results already exist for momentum accelerated optimization on Riemannian manifold, including both in continuous time (e.g., Alimisis et al., 2020) and in discrete time (e.g., Ahn and Sra, 2020). It is possible to formulate sampling as an optimization problem, where the convergence of a sampling dynamics/algorithm is characterized as conducting optimization in the infinite-dimensional space of probability densities (e.g., Vempala and Wibisono, 2019). However, when momentum is introduced to the manifold sampling problem, we have to be considering the convergence of densities defined on the tangent bundle of the manifold. This makes the analysis much more challenging. To the best of our knowledge, no convergence result exists for kinetic Langevin on curved spaces, in neither continuous or discrete time cases. See Sec. B.2 for more discussions of those difficulties.

This paper considers a special class of manifolds, known as the Lie groups. A Lie group is a differential manifold with group structure. Many widely used manifolds in machine learning indeed have Lie group structures; one example is the set of orthogonal matrices (with det=1), i.e. $\mathsf{SO}(n)$; see e.g., Lezcano-Casado and Martınez-Rubio (2019); Tao and Ohsawa (2020); Kong et al. (2023) for optimization on $\mathsf{SO}(n)$. However, fewer results have been established for sampling on Lie groups, especially with momentum. The interesting work by Arnaudon et al. (2019) might be the closest to ours in this regard, but the kinetic Langevin dynamics that we are proposing was not explicitly worked out, nor its discretization, and neither the uniqueness of invariant distribution or a convergence guarantee was provided, let alone convergence rate. In fact, similar to the general manifold case, we are unaware of prior construction of kinetic-Langevin-type samplers for Lie groups, and certainly not performance quantification.

1.1. A brief summary of main results

Our main contributions are:

1. We provide the first momentum version of Langevin-based algorithm for sampling on the curved spaces of Lie groups, with rigorous and quantitative analysis of the geometric ergodicities of both the sampling dynamics (i.e. in continuous time) on which it is based, and the sampling algorithm per se (i.e. in discrete time)

2. The exponential convergence on Lie groups is proved under weaker assumptions than considered in the literature for general manifolds, as we leverage their additional group structure. Only compactness of the manifold and smoothness of the log density are used. No (geodesic-)convexity or explicit isoperimetric inequalities are needed.

3. Our algorithm is fully implementable in the sense that it requires no implementation of Brownian motion on curved spaces. It is also computationally efficient as it is

based on explicit numerical discretization that preserves the manifold structure; that is, unlike commonly done, no extra projection back to the manifold (which can be computationally expensive) is needed. Note that our discretization is different from exponential integrators commonly used in Euclidean cases, and requires new analysis.

More specifically, consider sampling from a target distribution with density $\propto \rho(g)$. Letting $U = -\log\rho$ and calling it potential function ($U : \mathsf{G} \to \mathbb{R}$, where G is the Lie group), the sampling dynamics we construct is:

$$\begin{cases} \dot{g} = T_e \mathsf{L}_g \xi \\ \mathrm{d}\xi = -\gamma\xi \mathrm{d}t - T_g \mathsf{L}_{g^{-1}}(\nabla U(g))\mathrm{d}t + \sqrt{2\gamma}\mathrm{d}W_t \end{cases} \tag{1}$$

It admits the following invariant distribution

$$\nu_*(g,\xi) = \frac{1}{Z}\exp\left(-U(g) - \frac{1}{2}\langle\xi,\xi\rangle\right)\mathrm{d}g\mathrm{d}\xi \tag{2}$$

where $\mathrm{d}g$ is left Haar measure and $\mathrm{d}\xi$ is Lebesgue. Its g marginal is the target distribution.

The sampling algorithm we propose is:

Algorithm 1: Kinetic Langevin Monte Carlo (KLMC) sampler on Lie groups

---
Parameter : step size $h > 0$, friction $\gamma > 0$, number of iterations N
Initialization: $g_0 \in \mathsf{G}$, $\xi_0 = 0$
Output : A sample from $Z^{-1}\exp(-U(g))$
for $k = 0, \cdots, N-1$ **do**
$\quad$ $\xi_{k+1} = \exp(-\gamma h)\xi_k - \frac{1-\exp(-\gamma h)}{\gamma}T_g\mathsf{L}_{g^{-1}}\nabla U(g_k) + \mathcal{N}(0, 1 - \exp(-2\gamma h))$
$\quad$ $g_{k+1} = g_k\exp(h\xi_k)$
end
return g_N

The convergence guarantee for the sampling dynamics (eq.1; continuous time) is

Theorem 1 (Convergence of Kinetic Langevin dynamics on Lie groups (Informal version))
Suppose the Lie group G is compact, finite-dimensional and the potential function U is L-smooth, then we have the exponential convergence of sampling dynamics (1) to the Gibbs distribution ν_*, i.e.,

$$W_2(\nu_t, \nu_*) \leq C_\rho e^{-ct}W_\rho(\nu_0, \nu_*)$$

where ν_t is the joint density of $g(t)$ and $\xi(t)$. For notations and more details, see Thm. 8.

The nonasymptotic error bound for our sampler (Alg.1; discrete time) is

Theorem 2 (Convergence of Kinetic Langevin Sampler on Lie groups (Informal version))
Suppose the Lie group G is compact, finite-dimensional and the potential function U is L-smooth, then we have the exponential convergence of our KLMC algorithm to the Gibbs distribution, i.e.,

$$W_2(\tilde{\nu}_k, \nu_*) \leq C_\rho e^{-ckh}W_\rho(\nu_0, \nu_*) + \mathcal{O}(h^{1/2})$$

where $\tilde{\nu}_k$ is the density for the sampler at step k, ν_* is that target distribution in Eq. 2 and c is the contraction rate for the continuous dynamics. For notations W_ρ, C_ρ, a more explicit expression of $\mathcal{O}(h^{1/2})$, and more technical details, see Thm. 16.

2. Preliminaries

2.1. Lie group and Lie algebra

A **Lie group**, denoted as G, is a differentiable manifold with a group structure. A **Lie algebra** is a vector space with a bilinear, alternating binary operation that satisfies the Jacobi identity, known as Lie bracket. The tangent space at e (the identity element of the group) is a **Lie algebra**, denoted as $\mathfrak{g} := T_e\mathsf{G}$. The dimension of the Lie group G will be denoted by m. Lie groups considered when proving the convergence of kinetic Langevin dynamics and kinetic Langevin sampler will be assumed to satisfy the following:

Assumption 1 (General assumptions on geometry) *We assume the Lie group G is finite-dimensional, connected, and compact.*

The property we will use to handle momentum on Lie groups is called **left-trivialization**, which is an operation from the group structure and does not exist on a general manifold. Left group multiplication $\mathsf{L}_g : \hat{g} \to g\hat{g}$ is a smooth map from the Lie group to itself and its tangent map $T_{\hat{g}}\mathsf{L}_g : T_{\hat{g}}\mathsf{G} \to T_{g\hat{g}}\mathsf{G}$ is a one-to-one map. As a result, for any $g \in \mathsf{G}$, we can represent the vectors in $T_g\mathsf{G}$ by $T_e\mathsf{L}_g\xi$ for $\xi \in T_e\mathsf{G}$. This operation gives us an optimization dynamics on Lie groups in the next section.

2.2. Optimization dynamics based on left trivialization

Tao and Ohsawa (2020) proposed a Lie group optimizer based on constructing the following ODE, which performs optimization in continuous time on a general Lie group:

$$\begin{cases} \dot{g} = T_e\mathsf{L}_g\xi \\ \dot{\xi} = -\gamma\xi + \mathrm{ad}_\xi^* \xi - T_g\mathsf{L}_{g^{-1}}(\nabla U(g)) \end{cases} \tag{3}$$

ξ is the left-trivialized momentum (intuitively it could be thought of as angular momentum), and $T_e\mathsf{L}_g\xi$ is the momentum (intuitively, think it as $g\xi$, i.e. position times angular momentum gives momentum[1]). Potential $U : \mathsf{G} \to \mathbb{R}$ is the objective function of optimization. ∇U is the Riemannian gradient of the potential, and $T_g\mathsf{L}_{g^{-1}}(\nabla U(g))$ is its left-trivialization. This dynamics essentially models a damped mechanical system, where the total energy (sum of some kinetic energy term and potential energy U) is drained by the frictional forcing term $-\gamma\xi$, and U is minimized at $t \to \infty$. Indeed, this is how Tao and Ohsawa (2020) proved this ODE converges to a local minimum of U. The kinetic energy needs more discussion. In the Euclidean space, we have a global inner product, which is not true in curved spaces. In the Lie group case, we first define an inner product $\langle \cdot, \cdot \rangle$ on $\mathfrak{g}$, which is flat, and then move it around by the differential of left multiplication, i.e., the inner product at $T_g\mathsf{G}$ is for $\xi_1, \xi_2 \in T_g\mathsf{G}$, $\langle \xi_1, \xi_2 \rangle := \langle T_g\mathsf{L}_{g^{-1}}\xi_1, T_g\mathsf{L}_{g^{-1}}\xi_2 \rangle$. As a result, $\frac{1}{2}\langle \xi, \xi \rangle$ is the **kinetic** energy. γ provides dissipation to the total energy (the sum of kinetic energy and potential energy). In general, γ can be a positive function depending on time (e.g., NAG-C). However, we only consider the case γ as a constant for simplicity.

For curved space, an additional term $\mathrm{ad}_\xi^* \xi$ that vanishes in Euclidean space shows up in Eq. (3). It could be understood as a generalization of Coriolis force that accounts for

---

1. Technically, momentum is the dual of velocity, and these quantities should called velocity instead for rigor, but we will stick to the word 'momentum' by convention.

curved geometry and is needed for free motion, see Sec. C.3 for more discussion. The **adjoint operator** $\text{ad} : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ is defined by $\text{ad}_X Y := [X, Y]$. Its dual, known as the **coadjoint operator** $\text{ad}^* : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$, is given by $\langle \text{ad}_X^* Y, Z \rangle = \langle Y, \text{ad}_X Z \rangle, \forall Z \in \mathfrak{g}$.

2.3. Choice of inner product on $\mathfrak{g}$

The term $\text{ad}_\xi^* \xi$ in the optimization ODE (3) (and also the later sampling SDE Eq.4) is a quadratic term and it will make the numerical discretization that will be considered later difficult. Another more intrinsic drawback of this term is, it depends on the Riemannian metric, and indicates an inconsistency between the Riemannian structure and the group structure, i.e., the exponential map from the Riemannian structure is different from the exponential map from the group structure. See Sec. D.1 for more details. Fortunately, the following lemma shows a special metric on $\mathfrak{g}$ can be chosen to make the term $\text{ad}_\xi^* \xi$ vanish.

Lemma 3 (ad **skew-adjoint** (Milnor, 1976)) *Under Assumption 1, there exists an inner product on $\mathfrak{g}$ such that the operator* ad *is skew-adjoint, i.e.,* $\text{ad}_\xi^* = -\text{ad}_\xi$ *for any $\xi \in \mathfrak{g}$.*

Rmk. 26 in Sec. D.1 gives an explicit expression for this inner product. Under this inner product, $\text{ad}_\xi^* \xi = -\text{ad}_\xi \xi = [\xi, \xi] = 0$ because of the skew-symmetricity of the Lie-bracket. As a result, we can choose this inner product to ensure $\text{ad}_\xi^* \xi$ vanishes for any $\xi \in \mathfrak{g}$. This choice will be used throughout the rest of this paper. The diameter of G under such metric will be denoted as D.

Under the left-invariant metric induced by this inner product on $\mathfrak{g}$, the Lie group always has non-negative sectional curvature (Sec. D.2). Interestingly, different opinions exist about whether positive or negative curvatures help optimization or sampling. Positive curvatures help the quantification of discretization error via a modified cosine rule (Alimisis et al., 2020), while negative curvatures make it easier to have geodesically convex potential.

2.4. Assumptions on the potential function

After selecting our inner product on $\mathfrak{g}$ and using it to induce a left-invariant metric on G, the Riemannian structure on G enables us to define distance and gradient on G. With these, we can make the following smoothness assumption on the potential function U:

Assumption 2 (*L-smoothness*) *Under the left-invariant metric induced by the inner product in Lemma 3, there exist constants $L \in (0, \infty)$, s.t.*

$$\left\| T_g \mathsf{L}_{g^{-1}} \nabla U(g) - T_{\hat{g}} \mathsf{L}_{\hat{g}^{-1}} \nabla U(\hat{g}) \right\| \leq L d(g, \hat{g}) \quad \forall g, \hat{g} \in \mathsf{G}$$

Although here we compare gradients using left-trivialization, Lemma 28 shows the L-smoothness defined here is the same as the commonly used geodesic L-smoothness based on parallel transport (Def. 27), given the left-invariant metric in Lemma 3.

3. The SDE for sampling dynamics on Lie groups

In Euclidean space, when noise is added to momentum gradient descent, we have kinetic Langevin sampling SDE. Thanks to that our ξ is in a flat space, to obtain a Lie group generalization we add noise analogously to the optimization ODE Eq.(3) to obtain the following SDE on $\mathsf{G} \times \mathfrak{g}$,whose invariant distribution is the Gibbs distribution (Thm. 22):

$$\begin{cases} \dot{g} = T_e \mathsf{L}_g \xi \\ \mathrm{d}\xi = -\gamma \xi \mathrm{d}t + \text{ad}_\xi^* \xi \mathrm{d}t - T_g \mathsf{L}_{g^{-1}} (\nabla U(g)) \mathrm{d}t + \sqrt{2\gamma} \mathrm{d}W_t \end{cases} \tag{4}$$

Here dW_t means $\sum_i e_i dW_t^i$, with W^i being i.i.d. Brownian motion and $\{e_i\}_{i=1}^m$ being a set of orthonormal basis of $\mathfrak{g}$ under $\langle \cdot, \cdot \rangle$. The Brownian motion in both cases is not on a manifold but simply in a finite-dimensional vector space with an inner product, thus is well-defined.

The term $\mathrm{ad}_\xi^* \xi$ is again due to the curved space, see Sec.C.2. It ensures the invariant distribution (2) for any choice of inner product. See Thm. 22. However, if we use the inner product in Lemma 3, this term vanishes and the sampling SDE reduces to Eq.1.

4. Convergence of sampling dynamics in continuous time

We now prove the convergence of sampling SDE (4) under Wasserstein-2 distance. To do so, we first construct a coupling scheme by mixing synchronous coupling with reflection coupling. We also design a semi-distance. Using semi-martingale decomposition, we then prove the contractivity under this semi-metric, which induces convergence to the invariant distribution. Finally, this will be used to obtain convergence in W_2. The coupling scheme is inspired by Eberle et al. (2019), but both our semi-distance and detailed coupling scheme are different, which are specially designed to handle the curved space and to better utilize the compactness of the Lie group.

4.1. Construction of coupling

Coupling is a powerful probabilistic technique for, e.g., studying the convergence of a diffusion process. An easy example is synchronous coupling, i.e. consider $dx = b(x)dt + dW_t$ and $dy = b(y)dt + dW_t$ with the same noise and y initialized at the invariant distribution; when the drift term b provides contractivity, one can show x converges to y and hence the invariant distribution. For Euclidean Langevin SDE, it works, for example, when $b = -\nabla U$ with a strongly convex potential U. However, when contractivity from the drift term is not enough or noise is more complicated, we need more advanced coupling techniques:

Specifically, we consider (g_t, ξ_t) and $(\hat{g}_t, \hat{\xi}_t)$, both evolving in law as (4), given by:

$$\begin{cases} dg_t = T_e \mathsf{L}_{g_t} \xi_t dt \\ d\xi_t = -\gamma \xi_t dt - T_{g_t} \mathsf{L}_{g_t^{-1}}(\nabla U(g_t))dt + \sqrt{2\gamma}\, \mathrm{rc}(Z_t, \mu_t)dW_t^{\mathrm{rc}} + \sqrt{2\gamma}\, \mathrm{sc}(Z_t, \mu_t)dW_t^{\mathrm{sc}} \end{cases} \tag{5}$$

$$\begin{cases} d\hat{g}_t = T_e \mathsf{L}_{\hat{g}_t} \hat{\xi}_t dt \\ d\hat{\xi}_t = -\gamma \hat{\xi}_t dt - T_{\hat{g}_t} \mathsf{L}_{\hat{g}_t^{-1}}(\nabla U(\hat{g}_t))dt + \sqrt{2\gamma}(I_d - 2e_t e_t^\top)\, \mathrm{rc}(Z_t, \mu_t)dW_t^{\mathrm{rc}} + \sqrt{2\gamma}\, \mathrm{sc}(Z_t, \mu_t)dW_t^{\mathrm{sc}} \end{cases}$$

where $Z_t := \log \hat{g}_t^{-1} g_t$, $\mu_t := \xi_t - \hat{\xi}_t$ and $Q_t := Z_t + \frac{1}{\gamma}\mu_t$. $e_t = Q_t/\|Q_t\|$ if $Q_t \neq 0$, otherwise $e_t = 0$.

In all cases, we use log to define the inverse of the exponential map corresponding to the minimum geodesic (any one will do when the minimum geodesic is not unique). Moreover, $\mathrm{rc}, \mathrm{sc} : \mathbb{R}^{2m} \to [0, 1]$ are Lipschitz continuous functions such that $\mathrm{rc}^2 + \mathrm{sc}^2 \equiv 1$, and

$$\mathrm{rc}(z, \mu) = \begin{cases} 0, & \text{if } z + \gamma^{-1}\mu = 0 \text{ or } \|\mu\| \geq \gamma R + \epsilon, \\ 1, & \text{if } \left\| z + \gamma^{-1}\mu \right\| \geq \epsilon \text{ and } \|\mu\| \leq \gamma R \end{cases} \tag{6}$$

The parameters $R \in (0, \infty)$ will be chosen later in Sec. 4.3. ϵ is a fixed positive constant, and will go to 0 eventually (See the proof for Thm. 7).

The intuition of our design of coupling is illustrated in Fig. 1. The nonlinearly transformed phase space is partitioned into four regions where different couplings are used. More precisely -

Region (III) is where two momenta are very different. Their difference is so large that it is the main difference between (g, ξ) and $(\hat{g}, \hat{\xi})$. In this case, friction $-\gamma \xi$ provides contractility and synchronous coupling is sufficient for bringing the two copies together.

Region (II) is the 'good' part, which means the two points (g, ξ) and $(\hat{g}, \hat{\xi})$ will eventually stop at the same point under friction if the potential does nothing. Here, the momenta of the two copies are pulling their position variables together, thus providing contractivity. Synchronous coupling is enough for this region too.

Region (I) is the tricky part. None of the two ways above works, and we use reflection coupling instead to obtain contractivity. An intuition is the following: when reflection coupling is used, the difference between trivialized momenta is no longer differentiable

Figure 1: Partition of phase space using different couplings

in time but a diffusion process. Combined with the concavity of f in the semi-distance, which will be introduce in Sec. 4.2, a negative Ito correction term shows up in the adapted finite-variation process of a semi-martingale decomposition of the semi-distance, which provides an extra negative drift and leads to contractivity if the parameters are chosen carefully. More specifically, the helping term we are referring to is $4\gamma^{-1} \mathrm{rc}(Z, \mu)^2 f''(r)$ in K (defined in Sec. F.1) in the semi-martingale decomposition of $e^{ct} \rho_t$ (Lemma 6).

Region (IV) is the transitional region between synchronous coupling and reflection coupling, colored in gray in Fig. 1. Its has width ϵ and will eventually vanish when $\epsilon \to 0$.
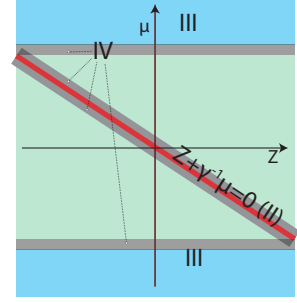
4.2. Design of semi-distance

Besides designing the coupling, we also need a quantification of how far two points are on $\mathsf{G} \times \mathfrak{g}$. One of our innovations is, we do not necessarily need a distance but only a semi-distance, i.e., no triangle inequality. We will construct the semi-distance carefully so that the coupling leads to contractivity. More precisely, let the semi-distance ρ be given by

$$\rho((g, \xi), (\hat{g}, \hat{\xi})) := f(r((g, \xi), (\hat{g}, \hat{\xi})))G(\xi, \hat{\xi}), \quad \forall g, \hat{g} \in \mathsf{G}, \forall \xi, \hat{\xi} \in \mathfrak{g} \tag{7}$$

where $f : [0, \infty) \to [0, \infty)$ is a continuous, non-decreasing, concave function satisfing $f(0) = 0$, $f'_+(0) = 1$, $f'_-(R) > 0$ [2] and f is constant on $[R, \infty)$. r and G are defined as

$$r((g, \xi), (\hat{g}, \hat{\xi})) := \alpha d(g, \hat{g}) + \left\| \log \hat{g}^{-1} g + \gamma^{-1}(\xi - \hat{\xi}) \right\| \tag{8}$$

$$G(\xi, \hat{\xi}) := 1 + \beta \left\| \hat{\xi} - \xi \right\|^2 \tag{9}$$

An explicit expression of f will be chosen later in Sec. F.3.1. The parameters $\alpha, \beta \in (0, \infty)$ will be given in Sec. F.3.2, whose values are chosen carefully together with other parameters (f and R) to achieve contractivity.

As shown, the semi-distance ρ has a complicated form, and the triangle inequality is sacrificed, which may lead to difficulties when analyzing numerical error from discretization later. But before diving into more details, let's provide some motivation for this complicated design. and how this design of semi-distance is related to our coupling.

ρ is the product of two parts, $f(r)$ and G. The function G is one plus the square of Euclidean distance between ξ and $\hat{\xi}$. It is designed for handling region (III), where friction

---

2. Here f_- and f_+ denote the left and right derivative of f, respectively.

ensures the decrease of G. Unlike Eberle et al. (2019), our G does not depend on position (g or $\hat{g}$). Besides the intuition we mentioned about friction, other reasons are: 1) In Eberle et al. (2019), G also provide contractivity when the positions g and $\hat{g}$ are far from each other utilizing strong dispativity under synchronous coupling. However, as discussed in Sec. B.3, in our case, we can not have convexity assumptions on the potential and including g and $\hat{g}$ in G does not help; 2) compactness ensures the distance between g and $\hat{g}$ is bounded, and in this case having G that only depends on the difference of momentum is enough to ensure ρ is stronger than d^2 (Lemma 4).

The other part of ρ is $f(r)$. It is by design concave because, like mentioned earlier, Ito's correction will give an additional term based on f'', and concavity of f will provide a negative sign and thus contractivity in region (I). Moreover, we make it first increasing and then constant starting from $r = R$ on. This way, $\min\{r \in \mathbb{R} : f(r) = \sup f\} = R$. Meanwhile, the bound for the synchronous coupling (III) is defined by $\|\xi - \hat{\xi}\| \geq \gamma R$. It is not a coincidence that they both scale with R; instead, it is because the coupling and semi-metric are carefully designed together: As mentioned earlier in Sec. 4.1, in region (III), the friction term is large enough and synchronous coupling suffices, i.e., when $\mu \geq \gamma R$, we have $r \geq R$. Therefore, the concavity of f in the interval of $r < R$ is sufficient for creating contractivity, and thus we simply set f to be constant on $[R, \infty)$.

Intuitively, r in Eq. (8) measures the 'distance' between points on $\mathsf{G} \times \mathfrak{g}$. The first parts $d(g, \hat{g})$ is directly the distance on the Lie group. However, we do not measure the difference between momenta directly, but by $\|\log \hat{g}^{-1} g + \gamma^{-1}(\xi - \hat{\xi})\|$, i.e. a twisted version with position distance also leaked in. The reason is similar to why we use region (I): we are measuring how far they will travel before eventually stopping due to friction without potential; this is a manifold generalization of the existing Euclidean treatment (e.g., Dalalyan and Riou-Durand, 2020; Eberle et al., 2019).

As a result, the design of ρ is a combination of functions that ensures contractivity in different regions. Later, we will carefully select parameters α and β to perfectly balance them. We will also discuss how to remedy the loss of triangle inequality and provide a substitute formula in Lemma 12 when we consider the discretization error of our SDE.

For now, we first show ρ can control the standard geodesic distance. The standard geodesic distance d in the product space $\mathsf{G} \times \mathfrak{g}$ is defined by

$$d((g, \xi), (\hat{g}, \hat{\xi})) = \sqrt{d^2(g, \hat{g}) + \|\hat{\xi} - \xi\|^2} \tag{10}$$

where d on the right-hand side is the distance on G given by the minimum geodesic length. Since both the distance on G and the distance on $\mathsf{G} \times \mathfrak{g}$ are derived from the inner product on $\mathfrak{g}$, we will use the same notation when there is no confusion. We have the following theorem showing d^2 on $\mathsf{G} \times \mathfrak{g}$ is controlled by ρ.

Lemma 4 (Control of d by ρ) *Define C_ρ as Eq. (48), we have*

$$C_\rho d^2((g, \xi), (\hat{g}, \hat{\xi})) \leq \rho((g, \xi), (\hat{g}, \hat{\xi})), \qquad \forall (g, \xi), (\hat{g}, \hat{\xi}) \in \mathsf{G} \times \mathfrak{g}$$

Here are more intuitions about ρ. For simplified notation, use d to denote $d((g, \xi), (\hat{g}, \hat{\xi}))$. When d is small, we have $r \sim d$ and $d^2 \lesssim d$. Moreover, we have f'_- lower bounded from 0, and $\rho \equiv f(r) \sim d$. However, when d is large, we have r is also large and $f(r) = f(R)$. In this

case, $\rho \sim f(R)d^2$. To summarize, ρ is similar to d when d is small but similar to d^2 when d is large. This is why d^2 but not d can be bounded by ρ in Lemma 4.

Remark 5 (Comparison with Eberle et al. (2019)) *The coupling and semi-distance used by Eberle et al. (2019) inspired our choices, but their version is not suitable for us. The first reason is that their G (Eq. 3.10 in Eberle et al. (2019)) is specially designed for the condition 'convex outside a ball', which is not available for Lie groups (Sec. B.3). Also, technical issues appear for their semi-distance:* $\mathrm{d}\log$ *(differential of logarithm) does not exist on a zero-measured set N (Eq.25) on Lie groups, and on a small neighbour of N, the operator norm of* $\mathrm{d}\log$ *can be unbounded, which leads to difficulties in calculation when using their sem-distance. However, our design of coupling and semi-distance utilizes the boundness of our Lie group. Consequently, in all our proof, what we only need for* $\mathrm{d}\log$ *is the properties in Cor. 31 and the fact that N is zero-measured is enough for our approach.*

4.3. Contractivity of sampling dynamics under W_ρ distance

We define the Wasserstein semi-distance between distributions on $\mathsf{G}\times\mathfrak{g}$ as W_ρ, i.e., $W_\rho(\nu_1,\nu_2) \coloneqq \inf_{\pi\in\Pi(\nu_1,\nu_2)} \int \rho d\pi$ where Π is the set of all distributions on $(\mathsf{G}\times\mathfrak{g})^2$ with marginal distributions ν_1 and ν_2. We hope to prove the contractivity of densities under W_ρ distance. The proof uses the following idea:

For any pair of points (g,ξ) and $(\hat{g},\hat{\xi})$ that are coupled together in the way stated in Sec. 4.1, we now construct a martingale decomposition bound of $e^{ct}\rho((g,\xi),(\hat{g},\hat{\xi}))$, where c is our target contraction rate that will be chosen later. More precisely, as will be shown by Lemma 6, $e^{ct}\rho((g,\xi),(\hat{g},\hat{\xi}))$ is decomposed as the sum of an adapted finite-variation process and a continuous local martingale, where the former is upper bounded by $\int_0^t e^{cs}K((g_s,\xi_s),(\hat{g}_s,\hat{\xi}_s))\mathrm{d}s$ with $K:(\mathsf{G}\times\mathfrak{g})^2\to\mathbb{R}$ defined later in Sec. F.1.

Lemma 6 *Let $c,\epsilon\in(0,\infty)$, and suppose that $f:[0,\infty)\to[0,\infty)$ is continuous, non-decreasing, concave, and C^2 except for finitely many points. Then*

$$e^{ct}\rho_t \le \rho_0 + \int_0^t e^{cs}K_s ds + M_t \quad \forall t\ge 0 \quad a.e., \tag{11}$$

where (M_t) is a continuous local martingale, and $K_s \coloneqq K((g_s,\xi_s),(\hat{g}_s,\hat{\xi}_s))$

By taking expectation, we only need $K((g_s,\xi_s),(\hat{g}_s,\hat{\xi}_s)) \le 0, a.e., \forall s$ when $\epsilon\to 0$, where (g_s,ξ_s) and $(\hat{g}_s,\hat{\xi}_s)$ are coupled as in Sec. 4.1 . This will infer that $\mathbb{E}e^{ct}\rho((g_t,\xi_t),(\hat{g}_t,\hat{\xi}_t))$ is non-increasing and further gives us the convergence rate under W_ρ distance. We summarize the conditions needed for $K((g_s,\xi_s),f(\hat{g}_s,\hat{\xi}_s)) \le 0, a.e.$ in Sec. F.2 and later choose all the parameters f, α, β, R and c in Sec. F.3 such that these conditions are met. The choice of γ is also given to establish an explicit order of convergence rate. This gives our contractivity result for sampling SDE Eq. (4) under semi-distance ρ:

Theorem 7 *For any probability distributions ν_0 and $\hat{\nu}_0$ on $\mathsf{G}\times\mathfrak{g}$ that is absolute continuous w.r.t. $\mathrm{d}g\mathrm{d}\xi$, if we evolve them by the sampling SDE (4), then, for some α, β, R, f, and $c>0$ (specified in Thm. 38), we have*

$$W_\rho(\nu_t,\hat{\nu}_t) \le e^{-ct}W_\rho(\nu_0,\hat{\nu}_0)$$

As we mentioned, W_ρ is only a semi-distance and lacks triangle inequality since ρ is only a semi-distance. However, what we only need for now is it controls W_2 distance:

4.4. Error bound for sampling SDE under W_2 distance

Upon choosing $\hat{\nu}_0$ in Thm.7 as the invariant distribution, obviously $\hat{\nu}_t = \hat{\nu}_0$ and Thm.7 thus quantifies the convergence speed of the sampling dynamics in W_ρ. Since W_ρ is what we invented and not a distance, we control d^2 by ρ, which infers W_ρ controls W_2^2 (Cor. 39), so that we can have the following theorem for convergence in a more standard distance:

Theorem 8 (Error of sampling SDE under W_2) *Suppose the initial condition $(g_0, \xi_0) \sim \nu_0$, where we assume ν_0 is absolute continuous w.r.t. $\mathrm{d}g\mathrm{d}\xi$. Denote by ν_* the Gibbs distribution Eq. (2) and by ν_t the distribution evolved by SDE Eq. (4). Then we have the W_2 distance between ν_t and ν_* is bounded by*

$$W_2(\nu_t, \nu_*) \le e^{-\frac{c}{2}t}\sqrt{C_\rho W_\rho(\nu_0, \nu_*)}$$

The reason why we need absolute continuity of the initial condition is because this gives us absolute continuity at any time t, which further enables us to ignore a bad set N where log is not differentiable. However, this condition will not lead to infeasibility of our discrete algorithm, which will be discussed later in Rmk. 36.

The contraction rate c is not explicitly expressed in the above theorem only because it is lengthy. A detailed expression is given by Eq. (42) in Lemma 38 and Eq. (47) gives an estimation of the order of c. How these results help choose γ is also discussed in Sec. F.4.

5. Convergence of sampling algorithm in discrete time

This section will first construct a sampler based on a time discretization of our sampling SDE. Thanks to an operator splitting technique, this discretization will render the iterations exactly satisfying the geometry of the curved space, which is a pleasant property referred to as structure-preservation (Thm. 9). Then we will establish its error bound in W_2 by: 1) quantifying local integration error in d (Thm. 10); 2) developing a modified triangle inequality for our semi-distance ρ (Lemma 12) and estimating sampling error propagation for the discrete sampler under $\mathbb{E}\rho$ (Thm. 13); 3) quantifying how local error accumulates to establish a global nonasymptotic error estimate under $\mathbb{E}\rho$ (Cor. 14), and then turning it into a nonasymptotic sampling error bound under W_ρ (Thm. 15); 4) bounding sampling error in W_2 (Thm. 16) by the fact that W_2^2 can be controlled by W_ρ (Cor. 39).

5.1. Sampler based on splitting discretization

We consider sampling SDE (4) with vanishing $\mathrm{ad}_\xi^* \xi$ (due to Lemma 3). To obtain a time discretization that respects the geometry, we write (4) as the sum of the following two SDEs, each of which can be solved explicitly, and alternatively evolve them to approximate the solution of (4), whose closed-form solution does not exist.

$$\begin{cases} \dot{g} = 0 \\ \mathrm{d}\xi = -\gamma\xi\mathrm{d}t - T_g\mathsf{L}_{g^{-1}}(\nabla U(g))\mathrm{d}t + \sqrt{2\gamma}\mathrm{d}W_t \end{cases} \quad (12) \qquad \begin{cases} \dot{g} = T_e\mathsf{L}_g\xi \\ \mathrm{d}\xi = 0 \end{cases} \quad (13)$$

To implement our algorithm, we embed the Lie group in an ambient Euclidean space, and the algorithm automatically keeps the iterates on the Lie group. Most algorithms in curved spaces similarly rely on this embedding, but they need extra work to correct the deviation

from manifold after each iteration, e.g., Cisse et al. (2017). In contrast, thanks to our specific splitting discretization, no matter how large the step size h is, our algorithm ensures g stays on the Lie group, and no artificial step that pulls the point back to the curved space is needed. Specifically, by first evolving Eq. (12) for time h and then Eq. (13) for time h, we obtain the following one step update:

$$
\begin{cases}
\tilde{g}_h = g_0 \exp(h\tilde{\xi}_h) \\
\tilde{\xi}_h = \exp(-\gamma h)\xi_0 - \frac{1-\exp(-\gamma h)}{\gamma} T_g \mathsf{L}_{g^{-1}} \nabla U(g_0) + \sqrt{2\gamma} \int_0^h \exp(-\gamma(h-s)) \mathrm{d}W_s
\end{cases}
\tag{14}
$$

Iterating this update gives Algorithm 1.

Its general structure preservation is summarized below; an example is given in Rmk.40.

Theorem 9 *The splitting discretization Eq. (14) is structure-preserving, i.e., for any step size h and any initial point (g_0, ξ_0), the iteration $(g_k, \xi_k), \forall k \geq 0$ has the property that g_k stays exactly on the Lie group and ξ_k stays exactly on the Lie algebra.*

We remark that the commonly used exponential integrator (see Eq.49) doesn't work here due to the nonlinear geometry. For example, the matrix-group-embedded g dynamics is $\dot{g} = g\xi$ where both g and ξ are matrices; because ξ is time-dependent, g admits no analytical solution. Existing tools for analyzing exponential-integrator-based samplers unfortunately have technical difficulties to be generalized here too (see Sec. H.1).

5.2. Quantification of discretization error

<u>Notation</u>: *starting from initial conditions (g_0, ξ_0) and $(\hat{g}_0, \hat{\xi}_0)$, we use (g_t, ξ_t) and $(\hat{g}_t, \hat{\xi}_t)$ to denote the exact solutions of two SDEs that are coupled together (5) correspondingly. $(\tilde{g}_h, \tilde{\xi}_h)$ is for one iteration by our sampler (i.e. splitting discretization (14)) from initial condition (g_0, ξ_0) with fixed step size h. $(\tilde{g}_k, \tilde{\xi}_k)$ is the result of k iterations.*

It is challenging to quantify discretization error directly in ρ, because ρ has a complicated expression and lacks triangle inequality. We bypass the difficulty by using the natural distance d on $\mathsf{G} \times \mathfrak{g}$ instead. The following theorem will quantify the one-step mean square error of our numerical scheme. The proof uses a new technique, namely to view the numerical solution after just one-step, with step size h, as the exact solution of some shadowing SDE at time h (Lemma 42), and then we only need to quantify the difference between the sampling SDE (4) and the shadow SDE (50).

Theorem 10 (Local numerical error) *Suppose the initial condition $(g_0, \xi_0) \sim \nu_0$ is absolute continuous w.r.t. $\mathrm{d}g\mathrm{d}\xi$. (g_t, ξ_t) is the solution of SDE (4) at time h and $(\tilde{g}_h, \tilde{\xi}_h)$ is the solution of the numerical discretization Eq. 14 with time step h. Then we have*

$$
\mathbb{E}d^2((g_h, \xi_h), (\tilde{g}_h, \tilde{\xi}_h)) \leq x(h) + y(h)
$$

where $x(t)$ and $y(t)$ are the solutions to an ODE given by Eq.52.

If one wants a more explicit bound, the following Lemma gives a less tight but more succinct one: Basically the local strong error is $3/2 = 1.5$ order.

Lemma 11 (Order of local error)

$$
\mathbb{E}d^2((g_h, \xi_h), (\tilde{g}_h, \tilde{\xi}_h)) = \mathcal{O}(h^3), \quad h \to 0
$$

i.e., $\exists h_0, A_{h_0}, s.t. \forall h \leq h_0, x(h) + y(h) \leq A_{h_0} h^3$ for x and y defined by Eq. 52.

11

Then, in the next subsection, we will combine this local error in d^2 and the contractivity of the continuous dynamics to establish the contractivity of our sampler.

5.3. Local error propagation

A standard technique for analyzing the sampling error of a sampler based on the discretization of a continuous dynamics is to transfer its infinite long time numerical integration error to a Wasserstein control of the difference between the continuous and discrete dynamics. To do so, one typically first seeks the contractivity of the continuous dynamics (Thm. 7), and then use that to both control the accumulation of local integration errors into global integration error, and bound the distance between the continuous dynamics and the target distribution so that the distance between the discrete dynamics and the target can also be bounded. For both tasks, triangle inequality is leveraged. However, as mentioned earlier in Sec. 4.2, ρ is only a semi-distance and does not satisfy triangle inequality, and we develop the following lemma as an alternation.

Lemma 12 (Modified triangle inequality for ρ) *For any points (g, ξ), $(\hat{g}, \hat{\xi})$ and $(\tilde{g}, \tilde{\xi})$,*

$$\rho((\hat{g}, \hat{\xi}), (\tilde{g}, \tilde{\xi})) \le \left[1 + A_1 d((g, \xi), (\tilde{g}, \tilde{\xi}))\right] \rho((\hat{g}, \hat{\xi}), (g, \xi)) + A_2 d^2((g, \xi), (\tilde{g}, \tilde{\xi}))$$

for some constants $A_1, A_2 > 0$ (their detailed expressions are given in Eq. 53).

Combining contractivity of continuous SDE (Thm.7), local discretization error (Thm.10) and modified triangle inequality for ρ (Lemma 12), we have the following result quantifying how error propagates after one step (more precisely, given two initial conditions, one evolving under the continuous dynamics for time h and the other iterated by the discrete algorithm (14) for one h-step, how the difference of the results changes from the initial difference):

Theorem 13 (Propagation of error under $\mathbb{E}\rho$) *Suppose $\mathrm{Law}(g_0, \xi_0)$ and $\mathrm{Law}(\hat{g}_0, \hat{\xi}_0)$ are absolute continuous w.r.t. $\mathrm{d}g\mathrm{d}\xi$. After one iteration from $(\tilde{g}_0, \tilde{\xi}_0) = (g_0, \xi_0)$ to $(\tilde{g}_h, \tilde{\xi}_h)$ by Eq.(14) with step size h, and exact evolution of $\hat{g}_0, \hat{\xi}_0$ by time h, we have, $\forall h > 0$,*

$$\mathbb{E}\rho((\hat{g}_h, \hat{\xi}_h), (\tilde{g}_h, \tilde{\xi}_h)) \le \exp(-ch)\mathbb{E}\rho((\hat{g}_0, \hat{\xi}_0), (g_0, \xi_0)) + E(h)$$

where $E(h)$ is given by

$$E(h) := \left(\frac{A_1}{2h^{\frac{3}{2}}} + A_2\right)(x(h) + y(h)) + \frac{A_1}{2} f(R)^2 (1 + 8C_2 + 16C_4) h^{\frac{3}{2}} \tag{15}$$

using the notation $x(h)$ and $y(h)$ from Lemma 10.

5.4. Mixing in W_ρ

By applying this local error propagation recurrently, we can obtain a global result:

Corollary 14 (Nonasymptotic error bound under $\mathbb{E}\rho$) *Under the same condition as Thm. 13, we have, for $k = 0, 1, \cdots$,*

$$\mathbb{E}\rho((\hat{g}_{kh}, \hat{\xi}_{kh}), (\tilde{g}_k, \tilde{\xi}_k)) \le e^{-ckh}\mathbb{E}\rho((\hat{g}_0, \hat{\xi}_0), (\tilde{g}_0, \tilde{\xi}_0)) + \frac{E(h)}{1 - \exp(-ch)}$$

The two semi-distances W_ρ and $\mathbb{E}\rho$ are closely related, and we can derive the following error bound in W_ρ, for measuring the difference between distributions.

Theorem 15 (Nonasymptotic error bound under W_ρ) *Given Assumption 1, 2 and the inner product on $\mathfrak{g}$ chosen as Lemma 3, if the initial condition $(g_0, \xi_0) \sim \nu_0$ satisfies $W_\rho(\nu_0, \nu_*) < \infty$ and ν_0 is absolute w.r.t. $\mathrm{d}g\mathrm{d}\xi$, then $\forall k = 1, 2, \ldots$, the density of scheme Eq. (14) starting from ν_0 has the following W_2 distance from the target distribution:*

$$W_\rho(\tilde{\nu}_k, \nu_*) \le e^{-ckh} W_\rho(\nu_0, \nu_*) + \frac{E(h)}{1 - \exp(-ch)},$$

where $E(h)$ is defined in Eq.15. Note this holds $\forall h > 0$, but $E(h)$ can grow exponentially.

5.5. Global sampling error in W_2

Via the global W_ρ error bound (Thm.15) and the property that d^2 is controlled by ρ (Lemma 4), we can also have a nonasymptotic control of the sampling error in a more common way of measurement, namely W_2, which is a true distance between distributions this time:

Theorem 16 (Nonasymptotic error bound under W_2) *Under the same assumption as Thm. 15, we have*

$$W_2(\tilde{\nu}_k, \nu_*) \le C_\rho \left(e^{-ckh} W_\rho(\nu_0, \nu_*) + \frac{E(h)}{1 - \exp(-ch)} \right)$$

The requirement that initial condition is absolute continuous w.r.t. $\mathrm{d}\mathfrak{g}\mathrm{d}\xi$ is only for technical reasons for the proof but not needed in implementation. See Rmk. 36 for details.

Remark 17 *Notice that $E(h)$ in Eq. (15) is of order $h^{\frac{3}{2}}$ and $1 - \exp(-ch)$ is of order h, which means the bias in the second term created by discretization error converges to 0 when step size is infinitely small, and the bias is asymptotically of order $h^{0.5}$.*

6. Numerical demonstration

Consider an example of sampling on the Lie group $\mathsf{SO}(n)$, which we embed in the matrix space $\mathbb{R}^{n \times n}$, i.e., $\mathsf{SO}(n) := \{X \in \mathbb{R}^{n \times n} : X^\top X = I, \det(X) = 1\}$ (Example 2). The inner product in Lemma 3 is given by $\langle A, B \rangle = \mathrm{tr}(A^\top B)$. Under this innner product, the left-trivialized gradient $T_g \mathsf{L}_{g^{-1}} \nabla U(g)$ can be calculated as follows: suppose $\mathsf{SO}(n)$ is represented by $n \times n$ matrices, and U is a real valued function defined on a subset of $\mathbb{R}^{n \times n}$. Denote its Euclidean gradient $\nabla_E U$, which can be calculated either by backward propagation or by closed-form solution. In this case, the Riemannian gradient $\nabla U(X)$ is given by $\nabla_E U(X) - X (\nabla_E U(X))^\top X$ and the left-trivialized gradient $T_g \mathsf{L}_{g^{-1}} \nabla U(g)$ is given by $X^{-1} \nabla U(X) = X^\top \nabla_E U(X) - (\nabla_E U(X))^\top X$.

Now we numerically experiment on a specific potential $U(X) = -10X_{1,1}^2$. We let $n = 10$, and the dimension of the group is thus 45. Note even though the potential might appear like a convex function, the target distribution is actually multimodal, and we recall again there is nonconstant geodesic convex function on this manifold (e.g., Sec.B.3).

The parameters for our sampler Algo. 1 are $\gamma = 1$ and $h = 0.1$. All the Markov chains are initialized by a fixed random generated orthogonal matrix. We focus on the variable $X[0,0]$ (the top left element of the matrix).

Fig. 2(a) compares the samples from our sampler to the ground truth samples generated by rejection sampling, showing our sampler is sampling the correct Gibbs distribution. Fig. 2(b) is another evidence that our sampler is dynamically sampling from a multimodal distribution. Exponential convergence of our sampler can be observed from Fig. 2(c), as theoretically proved in Thm. 8 and 16 despite of the multimodality. Fig. 2(d) and 2(e) visualize how the density evolves as our discrete process mixes and equilibrates.
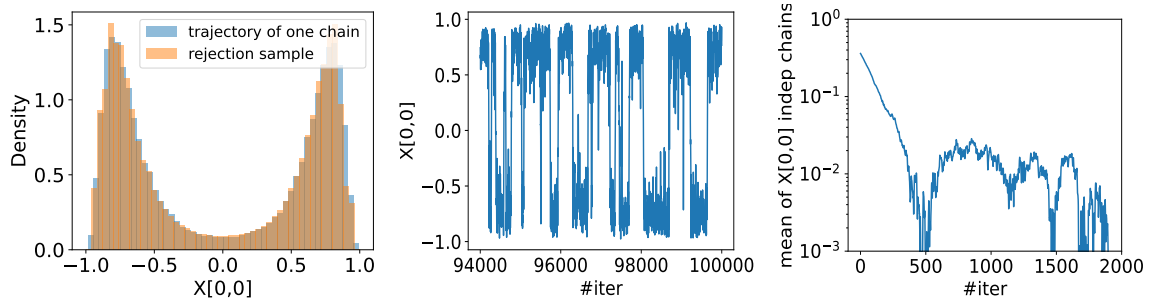
(a) Histogram of values collected along a single trajectory, compared to ground truth

(b) Trajectory of one chain

(c) Time evolution of the mean of an ensemble of chains, illustrating speed of convergence
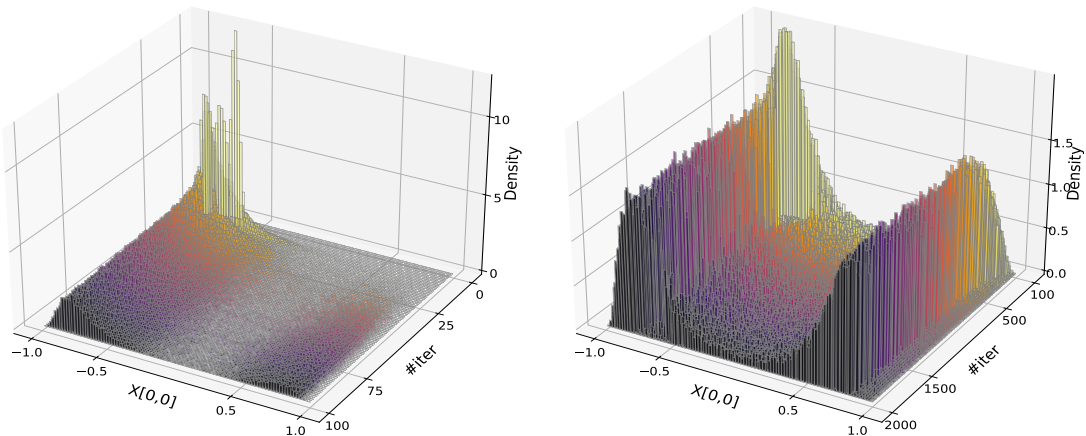
(d) Time evolution of the distribution of an ensemble of independent chains, prior to convergence (iteration #0-100)

(e) Time evolution of the distribution of an ensemble of independent chains, closer to convergence (iteration #100-2000)

Figure 2: Visualizing the exponential convergence of our Lie group sampler (Alg.1) for sampling a multimodal distribution on SO(10).

Acknowledgments

References

Ralph Abraham and Jerrold E Marsden. Foundations of mechanics. *Foundations of Mechanics*, 1978.

ID Ado. The representation of lie algebras by matrices. *Uspekhi Matematicheskikh Nauk*, 2 (6):159–173, 1947.

Kwangjun Ahn and Suvrit Sra. From nesterov's estimate sequence to riemannian acceleration. In *Conference on Learning Theory*, pages 84–118. PMLR, 2020.

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

Jason M Altschuler and Sinho Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. *arXiv preprint arXiv:2302.10249*, 2023.

Alexis Arnaudon, Alessandro Barp, and So Takao. Irreversible langevin mcmc on lie groups. In *Geometric Science of Information: 4th International Conference, GSI 2019, Toulouse, France, August 27–29, 2019, Proceedings 4*, pages 171–179. Springer, 2019.

Theodor Bröcker and Tammo Tom Dieck. *Representations of compact Lie groups*, volume 98. Springer Science & Business Media, 2013.

Xiang Cheng and Peter Bartlett. Convergence of langevin mcmc in kl-divergence. In *Algorithmic Learning Theory*, pages 186–211. PMLR, 2018.

Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped langevin mcmc: A non-asymptotic analysis. In *Conference on learning theory*, pages 300–323. PMLR, 2018.

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

Sinho Chewi. *Log-concave sampling*. draft, 2024. URL https://chewisinho.github.io/main.pdf.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International conference on machine learning*, pages 854–863. PMLR, 2017.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):651–676, 2017.

Arnak S Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic langevin diffusions. 2020.

Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of langevin monte carlo via convex optimization. *The Journal of Machine Learning Research*, 20(1):2666–2711, 2019.

EB Dynkin. Calculation of the coefficients in the campbell–hausdorff formula. *DYNKIN, EB Selected Papers of EB Dynkin with Commentary. Ed. by YUSHKEVICH, AA*, pages 31–35, 2000.

Andreas Eberle. Reflection couplings and contraction rates for diffusions. *Probability theory and related fields*, 166:851–886, 2016.

Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Couplings and quantitative contraction rates for langevin dynamics. 2019.

Murat A Erdogdu, Rasa Hosseinzadeh, and Shunshi Zhang. Convergence of langevin monte carlo in chi-squared and rényi divergence. In *International Conference on Artificial Intelligence and Statistics*, pages 8151–8175. PMLR, 2022.

Khashayar Gatmiry and Santosh S Vempala. Convergence of the riemannian langevin algorithm. *arXiv preprint arXiv:2204.10818*, 2022.

Nicolas Guigui and Xavier Pennec. A reduced parallel transport equation on lie groups with a left-invariant metric. In *Geometric Science of Information: 5th International Conference, GSI 2021, Paris, France, July 21–23, 2021, Proceedings 5*, pages 119–126. Springer, 2021.

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

Lingkai Kong and Molei Tao. Quantitative convergences of lie group momentum optimizers. *arXiv preprint arXiv:2405.20390*, 2024.

Lingkai Kong, Yuqing Wang, and Molei Tao. Momentum stiefel optimizer, with applications to suitably-orthogonal attention, and optimal transport. *ICLR*, 2023.

Mario Lezcano-Casado and David Martınez-Rubio. Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group. In *International Conference on Machine Learning*, pages 3794–3803. PMLR, 2019.

Ruilin Li, Hongyuan Zha, and Molei Tao. Sqrt(d) dimension dependence of langevin monte carlo. In *International Conference on Learning Representations*, 2021.

Yi-An Ma, Niladri S Chatterji, Xiang Cheng, Nicolas Flammarion, Peter L Bartlett, and Michael I Jordan. Is there an analog of nesterov acceleration for gradient-based mcmc? *Bernoulli*, 2021.

John Milnor. Curvatures of left invariant metrics on lie groups, 1976.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

André Schlichting. Poincaré and log–sobolev inequalities for mixtures. *Entropy*, 21(1):89, 2019.

Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. *Advances in Neural Information Processing Systems*, 32, 2019.

Molei Tao and Tomoki Ohsawa. Variational optimization on lie groups, with examples of leading (generalized) eigenvalue problems. In *International Conference on Artificial Intelligence and Statistics*, pages 4269–4280. PMLR, 2020.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

Xiao Wang, Qi Lei, and Ioannis Panageas. Fast convergence of langevin dynamics on manifold: Geodesics meet log-sobolev. *Advances in Neural Information Processing Systems*, 33:18894–18904, 2020.

Shing-Tung Yau. Non-existence of continuous convex functions on certain riemannian manifolds. *Mathematische Annalen*, 207:269–270, 1974.

Bo Yuan, Jiaojiao Fan, Yuqing Wang, Molei Tao, and Yongxin Chen. Markov chain monte carlo for gaussian: A linear control perspective. *IEEE Control Systems Letters*, 2023.

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *Conference on learning theory*, pages 1617–1638. PMLR, 2016.

Shunshi Zhang, Sinho Chewi, Mufan Li, Krishna Balasubramanian, and Murat A Erdogdu. Improved discretization analysis for underdamped langevin monte carlo. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 36–71. PMLR, 2023.

Category	Notation	Location	Description
Basic	G		Lie group
	$\mathfrak{g}$		Lie algebra
	g	Sec. 2.1	Element in Lie group
	ξ		Element in Lie algebra
	m		dimension of Lie groups /algebra
Group structure	ad	Sec. 2.2	Adjoint operator
	c_{ij}^k	Def. 20	Structural constant
	C	Eq. (21)	Operator norm of ad
	N	Eq. (25)	Set that $\mathrm{d}\log$ do not exist
Riemannian structure	ad^*	Sec. 2.2	Coadjoint operator
	Γ_{ij}^k	Sec. D.2	Christoffel symbol
	∇		Gradient or Levi-Civita connection
	d^3	Eq. (10)	distance on G or $\mathsf{G} \times \mathfrak{g}$
	D	Sec. 2.3	Diameter of G
Convergence	$(g_t, \xi_t), (\hat{g}_t, \hat{\xi}_t)$	Sec. 4.1	Coupled r.v. following sampling SDE
	$(\tilde{g}_t, \tilde{\xi}_t)$	Sec. 5.1	R.v. from discretization
	C_k	Lemma 46	Upper bound for $\mathbb{E}\|\xi\|^k$
	x, y	Eq. (52)	ODE quantifying local numerical error
	h		Step size for discretization
Others	ω	Eq. (18)	symplectic 2-form
	$\{e_i\}_{i=1}^m$		A set of orthonormal basis in $\mathfrak{g}$
	p and p^{-1}	Eq. (22), (23)	power series for $\mathrm{d}\log$ and $\mathrm{d}\exp$
	α, β, R, e_t, Z_t, Q_t	Sec. 4.1, 4.2	coupling and semi-distance
	K	Sec. F.1	drift term in decomposition of ρ

Table 1: A list of notations

Appendix A. Notation and map

Please see Table 1 for a list of notations and Fig. 3 for a graph showing the dependence of theorems.

---

3. We use d for distance and d for differential (please note the different fonts).
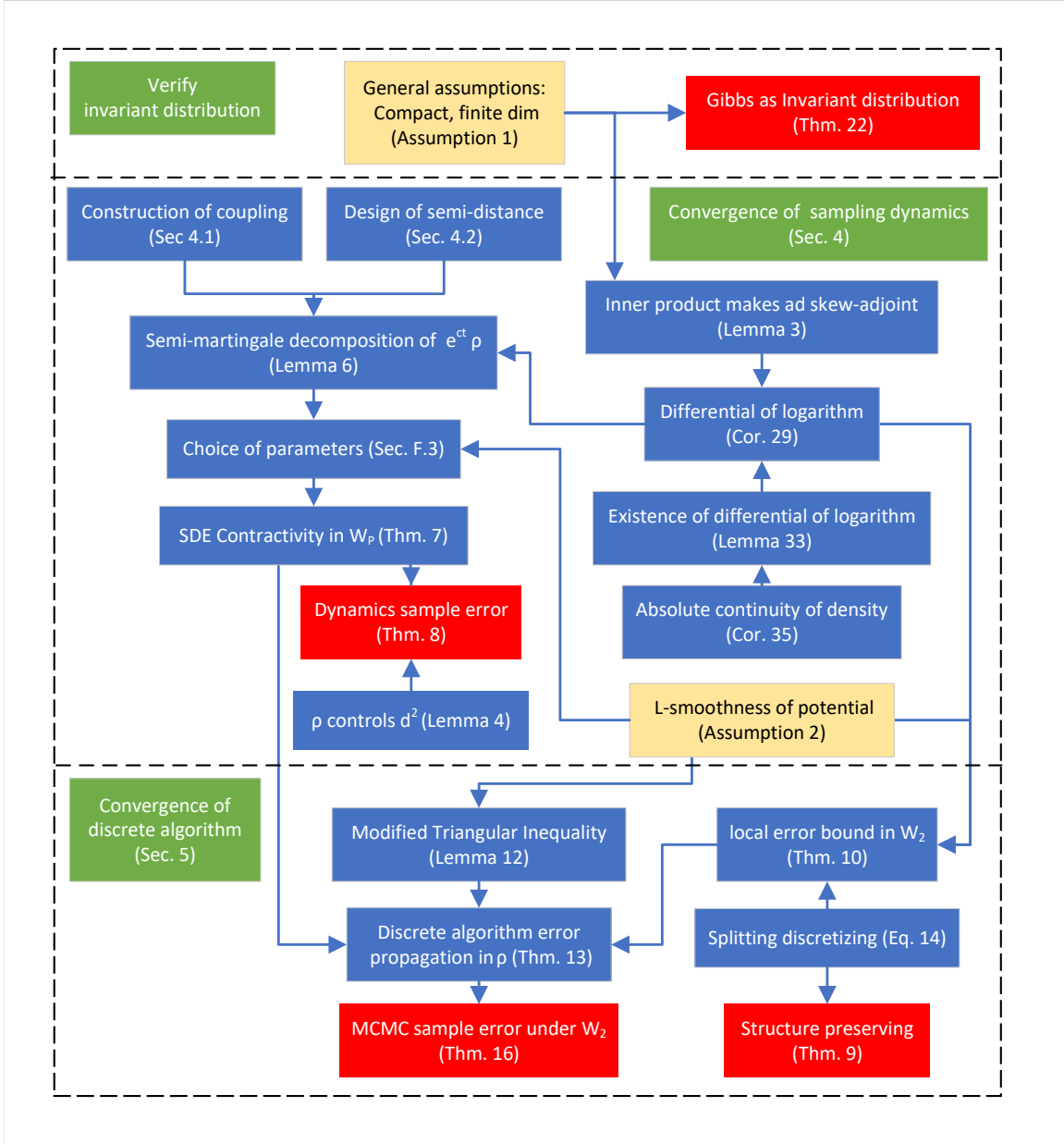
Figure 3: Diagram of the dependence of theorems

Appendix B. More discussion on related works

B.1. More discussion on Langevin sampling in Euclidean spaces

In order to analyze the convergence of Langevin-type sampling algorithms, at least two common ways of quantification have been used, based on different conditions and analysis techniques. 1) When convergence is quantified under Wasserstein distance, coupling methods provide useful tools. 2) When convergence is measured via f-divergence, a Lyapunov approach that quantifies convergence in the space of probability densities is often employed.

Roughly speaking, 2) requires weaker conditions than 1), but is more challenging to work with when momentum is involved. Here are some more details.

Wasserstein metric is a distance between measures. Coupling methods are usually used for proving the convergence rate due to the definition of Wasserstein distance. For over-damped Langevin, it is a standard proof that synchronous coupling can establish its contractivity under a strong convexity assumption of the potential. Kinetic Langevin is a little more complicated, but it was also known that when its friction coefficient is constant, contractivity still stands after a constant linear coordinate transformation, for both the continuous SDE and its discretization (e.g., Dalalyan and Riou-Durand, 2020; Cheng et al., 2018). Contractivity allows the convergence analysis of a continuous dynamics in Wasserstein distance to carry over to the analysis of its time-discretization, i.e. sampling algorithms (e.g., Li et al., 2021), and it was known that the sampling accuracy can be improved via good numerical discretization (e.g., Shen and Lee, 2019). In addition, coupling method can be generalized to potential function under weaker conditions, e.g., convexity outside a ball. For example, reflection coupling worked well for overdamped Langevin (Eberle, 2016), and the momentum case (i.e. kinetic Langevin) is later considered (Eberle et al., 2019). The clever but complicated design of coupling and semi-distance function employed in these approaches makes it nontrivial to obtain an explicit convergence rate for the discrete algorithm.

Another quantification of convergence is based on some f-divergence from one measure to another. For example, Langevin dynamics and its discretization of Langevin Monte Carlo can be shown convergent in various f-divergences under various isoperimetric inequality assumption on the target distribution, such as in KL under Logarithmic Sobolev Inequality (LSI) and in chi-square under Poincare inequality (PI) (e.g., Vempala and Wibisono, 2019; Erdogdu et al., 2022; Chewi, 2024). Since PI is weaker than LSI, and LSI is weaker than convexity outside a ball, in some sense convergence in f-divergence requires a weaker condition than that in Wasserstein distance[4]. The proof of the convergence by Vempala and Wibisono (2019), for instance, is based on splitting the Markov kernel into a deterministic part and a Brownian motion part. The deterministic part keeps the KL/Renyi divergence unchanged and only ensures the invariant distribution is correct. The Brownian motion part mollifies the density and leads to a monotonic decrease in the KL divergence between the current distribution and the target distribution. However, this approach for example is difficult to generalize to the momentum case (i.e. kinetic Langevin, frequently referred to as underdamped Langevin too) due to the degeneracy of noise. Unlike in the {Wasserstein distance + strong convexity} case where a constant linear change of coordinates helps recover contractivity, the analysis of convergence in f-divergence under isoperimetric inequalities is often a case-by-case study. For example, Ma et al. (2021) proved the convergence of KLMC in KL divergence under LSI assumption by adding a carefully designed cross term to the joint KL divergence. Zhang et al. (2023) provided additional tools that allow improved analysis of the convergence of KLMC (e.g., the smoothness assumption on Hessian in Ma et al. (2021) is no longer needed), and convergence in both KL and Renyi is provided. Altschuler and Chewi (2023) proposed an innovative technique based on Renyi divergence with Orlicz-Wasserstein shifts and used it to prove many great results, such as the conver-

---

4. Although better convergence rates can be established under stronger conditions (e.g. Dalalyan, 2017; Cheng and Bartlett, 2018; Durmus et al., 2019; Li et al., 2021).

gence rate with improved dimensional dependence of Metropolis-adjusted KLMC, under a variety of metric including total variance, Chi-square and KL divergence and W_2 distance with the assumption that the target distribution satisfies either LSI or PI.

B.2. When momentum meets curved spaces

Both curved space and momentum (which renders the underlying Markov process for sampling no longer reversible) lead to extra difficulties. Quantifying numerical error in curved spaces is much harder compared to the cases in flat spaces (e.g., Gatmiry and Vempala, 2022; Cheng et al., 2022), and momentum leads to the degeneracy of noise, which means extra techniques are needed (e.g., Dalalyan and Riou-Durand, 2020; Cheng and Bartlett, 2018; Shen and Lee, 2019; Ma et al., 2021; Zhang et al., 2023; Yuan et al., 2023; Altschuler and Chewi, 2023). However, when both curved space and momentum show up together, there is an extra difficulty: when studying kinetic Langevin, we are considering the convergence of measures on the phase space, i.e., the product space of position and momentum. When the space is curved, momentum is a vector in the tangent space[5] of the current position, and the phase space becomes the tangent bundle. Convergence analysis on the tangent bundle is challenging, due to different reasons in the f-divergence case and the Wasserstein distance case, and we will discuss them separately now.

When quantifying convergence in f-divergence, we are considering the ratio between the current density and the density of the invariant distribution, and the Fokker-Planck equation governs the density evolution, which involves the (spatial) gradient of density. Therefore, we can not bypass the need of calculating the gradient on the tangent bundle. However, a Riemannian metric is required to calculate the gradient, and in this case, we need a Riemannian metric for the entire tangent bundle instead of just the curved space. Although one can induce a Riemannian metric of the tangent bundle from the Riemannian metric of the manifold (e.g., Sasaki metric), how to apply it to construct and analyze a kinetic-type Langevin dynamics or its discretization is an open problem.

When quantifying convergence in Wasserstein distance, we need to compare how far two points are in the phase space, which is typically then used as the cost function to induce a Wasserstein distance. In the manifold case, one can compare momenta (i.e. vectors in tangent space) via parallel transport, which however could be path-dependent, i.e., different paths connecting the two points give different linear isomorphisms between tangent spaces. In other words, a uniform way to compare momenta at different positions does not exist. This difficulty is not severe in optimization, since we only consider a trajectory in that case. A natural choice is to move the momentum along the position trajectory (e.g., Ahn and Sra (2020)). However, in sampling tasks, the position as a random variable takes values over the whole manifold, and as a result, we need a uniform way to compare momenta, and so far we do not know how to do so. To the best of our knowledge, there is no known result that extends kinetic Langevin dynamics or KLMC to the manifold case, let alone any analysis of the convergence.

All those difficulties due to curved spaces and momentum still need to be addressed when we construct and analyze kinetic Langevin Monte Carlo on Lie groups. What further

---

5. Rigorously speaking, momentum should be in the cotangent space while velocity is in the tangent space, but we will follow the convention and not distinguish velocity and momentum.

complicates the problem is, the additional group structure is not free lunch. For example, here is one chain of complications: 1) Lie groups with left-invariant metric and sectional curvature negative everywhere are very limited (Milnor, 1976). 2) On many Lie groups, there exists two points such that there are more than one geodesics connecting them. This further implies: 3) (geodesic) strong convexity is too strong for many Lie groups as there is no nontrivial convex function along a closed geodesic. As a result, we do not want to make assumptions about the potential function other than smoothness. See Sec. B.3 for more details. However, despite all those difficulties from Lie groups, we can enjoy a useful technique, namely (left) trivialization, to establish a uniform way of comparing momenta in the tangent spaces at different points. Trivialization provides an alternative tool for bijection between tangent spaces, by related both to a fixed linear space known as the Lie algebra, and plays a key role in this paper. We will build the sampling dynamics by trivialized momentum and then have our structure-preserving discretization based on it. Our proofs will also heavily rely on trivialization.

B.3. Discussion on commonly used conditions for the potential function

As mentioned earlier, normally some assumptions are required for the potential function. L-smoothness is the most commonly assumed one, and our work also assumes it to compare the gradients of potential at different positions. Meanwhile, in most existing results, additional assumptions are used to ensure geometric ergodicity. More precisely -

In Euclidean space, when U is strongly convex, one can prove the convergence of kinetic Langevin by showing the contractivity of dynamics after a constant coordinate transformation (Dalalyan and Riou-Durand, 2020). It is unclear to us how to do the coordinate transformation in curved spaces, because it can no longer be constant due to the nonlinearity of space, but then additional challenges arise (e.g., it no longer induces a metric as essentially done in (Dalalyan and Riou-Durand, 2020)). In fact, even discussing what happens under convexity is vacuous on Lie groups because there is no nontrivial geodesically convex function on compact Lie group (Yau, 1974). An intuition for this is, any convex function on a closed geodesic must be constant.

One relaxation of strong convexity is distant-dissipativity (Cheng et al., 2022). In flat space, it is called strong-dissipativity (Eberle, 2016; Erdogdu et al., 2022), defined by

$$\langle \nabla U(x) - \nabla U(y), x - y \rangle \geq m_1 \|x - y\|^2, \forall \|x - y\| \geq R_1 \quad \text{for some} \quad m_1, R_1 > 0 \qquad (16)$$

also known as 'strongly convexity outside a ball'. Although such condition is widely used and helpful in allowing us to bypass nonconvexity by containing it inside a ball, it is however still a rather strong assumption. The reason is, a function with distant-dissipativity is still convex at a large scale, but no nontrivial (i.e. nonconstant) geodesically convex function exists on manifolds with closed geodesic. Fig.4 illustrates why this is the case.

There are weaker conditions, such as log-Soblev inequality (LSI) and Poincare inequality (PI), but it is unclear how to use them together with coupling methods for manifolds.

To summarize, the commonly used assumptions on the potential function are not necessarily suitable for our compact Lie group case, and this is why we only make the L-smooth assumption on the potential function.
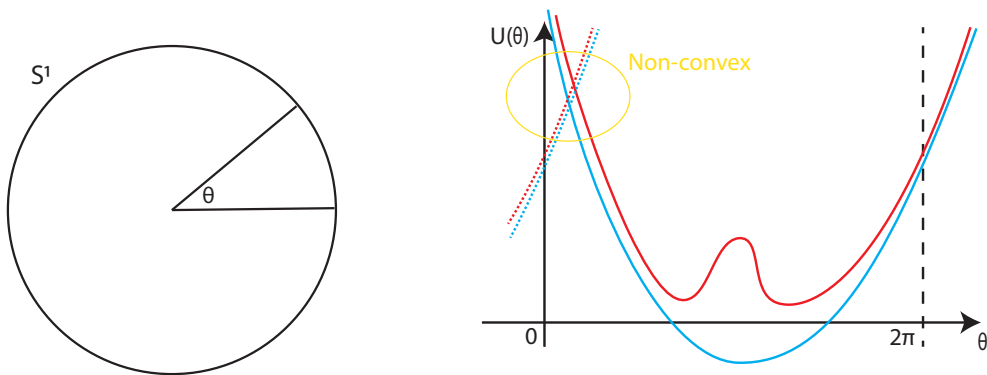
(a) Parameterization of $\mathsf{SO}(2) \cong S^1$ (b) Red solid: a function on $\mathsf{SO}(2)$ that might be mistaken as convex outside a ball; Blue solid: its 'convexification' which is not convex; Dashed lines: what periodicity and convexity would require, which create inconsistency.

Figure 4: An illustration of the non-existence of strongly convex or distant-dissipative function on a manifold with closed geodesics. The intuition is a convex function (if any) would grow in both sides, but along a loop, increasing θ eventually leads to θ reset to 0 and increasing it again, which means the function needs to be both increasing and decreasing as θ grows (e.g., the yellow region), which is a contradiction.

Appendix C. More details about Sec. 2 and 3

C.1. Examples for Lie groups

In this section, we give some examples of Lie groups. The most well-known Lie group is the Matrix group:

Example 1 (General Linear group) *Denote **general linear group** by $\mathsf{GL}(n, \mathbb{R}) := \{X \in \mathbb{R}^{n \times n} : \det X \neq 0\}$ with the group multiplication defined by matrix multiplication. The corresponding Lie algebra $\mathfrak{gl}(n, \mathbb{R}) := \mathbb{R}^{n \times n}$ has the matrix commutator: $[A, B] := AB - BA$ as its Lie bracket. Consequently, $T_e L_g \xi$ is $g\xi$ (matrix multiplication of two n-by-n matrices) and the exponential map is the matrix exponential.*

An natural choice of the inner product on $\mathfrak{gl}$ is $\langle A, B \rangle = \mathrm{tr}(A^\top B)$. Under the left-invariant metric induced by this inner product, $\mathrm{ad}_A^* B = BA^\top - A^\top B$. Note that the operator ad is not skew-adjoint in this case. In fact, the inner product that makes ad^* skew-adjoint does not exist. A necessary and sufficient condition for the existence of such inner product is provided in Milnor (1976).

The definition of general Lie groups as manifolds in Sec. 2.1 may sound abstract, but in fact, the following remark shows all compact Lie groups can be viewed as a subgroup of a matrix group in Example 1.

Remark 18 (Matrix representation of Lie groups/Lie algebra) *Peter–Weyl theorem implies that every compact Lie group is a closed subgroup of $\mathsf{GL}(n, \mathbb{R})$ for some n. Ado's*

theorem (*Ado, 1947*) *states that every finite-dimensional Lie algebra on* $\mathbb{R}$ *can be viewed as a Lie subalgebra of* $\mathfrak{gl}(n,\mathbb{R})$ *for some* n.

Although the general linear group is not compact making our assumption 1 fails, many of its Lie subgroups are compact. A nice example is the following:

Example 2 (Special Orthogonal group $\mathsf{SO}(n)$**)** *The special orthogonal group is defined as*

$$\mathsf{SO}(n) := \left\{ X \in \mathbb{R}^{n \times n} : X^\top X = I, \det(X) = 1 \right\}$$

The corresponding Lie algebra is $\mathfrak{so}(n) := \{\xi \in \mathbb{R}^{n \times n} : \xi^\top + \xi = 0\}$. *Since it is a Lie subgroup, the group operation, left translation, and exponential map are the same as general linear group (Example 1). The inner product that makes* ad *skew-adjoint in Lemma 3 is given by* $\langle A, B \rangle := \operatorname{tr}(A^\top B)$.

We also give an explicit example such that an inner product that makes ad skew adjoint (Lemma 3) may not exist when assumption 1 does not hold:

Example 3 (Lie group with constant negative curvature) *We consider a 2-dim Lie group whose Lie algebra has basis* e_1 *and* e_2. *Define a linear operator as* $l : \mathsf{G} \to \mathbb{R}$ *by* $l(xe_1 + ye_2) := ax + by, \forall x, y$ *for some constants* $a, b \neq 0$ *and the Lie bracket as* $[e_1, e_2] = l(e_1)e_2 - l(e_2)e_1$. *Example 1.7 in* Milnor (1976) *shows this Lie group has strictly negative sectional curvature for any left-invariant metric.*

We can explicitly express this example as a Lie subgroup of the general linear group using adjoint representation. Denote $\mathsf{G} := \{xe_1 + ye_2 : x, y \in \mathbb{R}\} \subset \mathsf{GL}(2, \mathbb{R})$ *where the basis are defined as* $e_1 = \begin{pmatrix} 0 & 0 \\ -b & a \end{pmatrix}$ *and* $e_2 = \begin{pmatrix} b & -a \\ 0 & 0 \end{pmatrix}$. *Together with Eq. 20, this tells us the inner product in Lemma 3 that makes* ad *skew-adjoint does not exist.*

The flat Euclidean space is a trivial Lie group and our sampling dynamics recover kinetic Langevin dynamics in the Euclidean space.

Example 4 (Lie group structure for Euclidean spaces) *Euclidean space is a trivial (commutative) Lie group, i.e., when* $\mathsf{G} = \mathbb{R}^n$ *with the group multiplication defined as* $g_1 g_2 := g_1 + g_2$ *(vector summation). The corresponding Lie algebra is* $\mathfrak{g} = \mathbb{R}^n$ *with vanish Lie bracket. The exponential map is the identity map on* $\mathbb{R}^n$.

The left-trivialization in the Euclidean space is the identity map, and Eq. (4) in Euclidean spaces becomes

$$\begin{cases} \dot{q} = p \\ \mathrm{d}p = -\gamma p \mathrm{d}t - \nabla U(q) \mathrm{d}t + \sqrt{2\gamma} \mathrm{d}W \end{cases} \tag{17}$$

which is the well-known kinetic Langevin dynamics.

Although many elegant results have been established for the Euclidean case (Sec. B.1, especially Eberle et al. (2019)), our analysis only focuses on compact Lie groups and cannot be applied to the Euclidean sampler (despite that our algorithm does).

C.2. Symplectic structure on Lie groups

The tangent bundle of a Lie group $T\mathsf{G} \cong \mathsf{G} \times \mathfrak{g}$ has a natural symplectic structure (Abraham and Marsden, 1978, Prop. 4.4.1). More specifically, at $(g, \mu) \in \mathsf{G} \times \mathfrak{g}$, the symplectic 2-form is given by

$$\omega_{(g,\mu)}((v, \alpha), (w, \beta)) = \langle \beta, T_g\mathsf{L}_{g^{-1}}v \rangle - \langle \alpha, T_g\mathsf{L}_{g^{-1}}w \rangle + \langle \mu, [T_g\mathsf{L}_{g^{-1}}v, T_g\mathsf{L}_{g^{-1}}w] \rangle \qquad (18)$$

for any $(v, \alpha), (w, \beta) \in \mathsf{G} \times T_g\mathsf{G}$.

Given the symplectic structure above, we can provide the optimization dynamics with a mechanic view. Setting the Hamiltonian $H : G \times \mathfrak{g} \to \mathbb{R}$ as $H(g, \xi) := U(g) + \frac{1}{2}\langle \xi, \xi \rangle$ gives a Hamiltonian field X_H, defined as the unique vector field on $\mathsf{G} \times \mathfrak{g}$ satisfying $\mathrm{d}H = \omega(X_H, \cdot)$. The Hamiltonian flow $\frac{\mathrm{d}}{\mathrm{d}t}(g_t, \xi_t) = X_H$ is exactly Eq. (3) with $\gamma = 0$. The term $\mathrm{ad}_\xi \xi$ in Eq. (3), which does not show in flat Euclidean spaces, comes from the third term in the symplectic 2-form in Eq. (18).

On a symplectic manifold, there is always a natural volume form $\omega^m := \omega \wedge \cdots \wedge \omega$, which provides a base measure. Arnaudon et al. (2019) uses ω^m as the base measure distribution when proving the invariant distribution. How is this symplecticity-based measure ω^m related to the group-structure-based measure $\mathrm{d}g\mathrm{d}\xi$ in Thm.22? They are identical, as proved in Thm. 19. Note that this theorem do not depend our special choice of inner product that makes ad skew-adjoint in Lemma 3.

Theorem 19 (Equivalence of base measure) *The product measure on $\mathsf{G} \times \mathfrak{g}$, i.e., the product of left Haar measure on G and Lebesgue measure on $\mathfrak{g}$, is identical to the measure induced by the volume form ω^m up to a constant.*

Proof [Proof of Thm. 19] The outline of this proof is: we will choose $2m$ vector fields on the tangent space of $2m$-dim manifold $\mathsf{G} \times \mathfrak{g}$, and we calculate the ratio of the two volume forms corresponding to those two different vector fields when applying to the $2m$ vector fields. We will prove the ratio is constant, showing the two measures are identical up to a constant.

We fix the following $2m$ vector fields, $\{(T_e\mathsf{L}_ge_i, 0)\}_{i=1,\ldots,m}$ and $\{(0, e_j)\}_{j=1,\ldots,m}$, where vector fields $T_e\mathsf{L}_ge_i$ on G are left invariant vector fields generated by e_i. In the following, we prove that $\omega^m_{(g,\mu)}((T_e\mathsf{L}_ge_1, 0), \ldots, (T_e\mathsf{L}_ge_m, 0), (0, e_1), \ldots, (0, e_m))$ is constant.

By the expression of ω in Eq. (18),

$$\omega_{(g,\mu)}((T_e\mathsf{L}_ge_i, 0), (T_e\mathsf{L}_ge_j, 0)) = \langle [e_i, e_j], \mu \rangle$$
$$\omega_{(g,\mu)}((T_e\mathsf{L}_ge_i, 0), (0, e_j)) = \delta_{ij}$$
$$\omega_{(g,\mu)}((0, e_i), (T_e\mathsf{L}_ge_j, 0)) = -\delta_{ij}$$
$$\omega_{(g,\mu)}((0, e_i), (0, e_j)) = 0$$

The definition of exterior product gives

$$\omega^m_{(g,\mu)}((T_e\mathsf{L}_ge_1, 0), \ldots, (T_e\mathsf{L}_ge_m, 0), (0, e_1), \ldots, (0, e_m))$$
$$= \frac{1}{(2m)!} \sum_{\sigma \in S_{2m}} \mathrm{sgn}(\sigma)\omega^m_{(g,\mu)}\sigma((T_e\mathsf{L}_ge_1, 0), \ldots, (T_e\mathsf{L}_ge_m, 0), (0, e_1), \ldots, (0, e_m))$$

By the fact that $\omega_{(g,\mu)}((0,e_i),(0,e_j))$ and $\omega_{(g,\mu)}((T_e\mathsf{L}_ge_i,0),(T_e\mathsf{L}_ge_j,0))+\omega_{(g,\mu)}((T_e\mathsf{L}_ge_j,0),(T_e\mathsf{L}_ge_i,0))$ vanishes, we have that all the non-vanish terms must have the form $\omega_{(g,\mu)}((T_e\mathsf{L}_ge_i,0),(0,e_i))$, and

$$\omega_{(g,\mu)}^m\left((T_e\mathsf{L}_ge_1,0),\ldots,(T_e\mathsf{L}_ge_m,0),(0,e_1),\ldots,(0,e_m)\right)$$

$$= \frac{1}{(2m)!}\sum_{\sigma\in S_m}2^m\prod_{i=1,\ldots,m}\omega_{(g,\mu)}\sigma\left((T_e\mathsf{L}_ge_i,0),(0,e_i)\right)$$

$$= \frac{m!2^m}{(2m)!}\prod_{i=1,\ldots,m}\omega_{(g,\mu)}\sigma\left((T_e\mathsf{L}_ge_i,0),(0,e_i)\right)$$

$$= \frac{m!2^m}{(2m)!}$$

Since both ω^m and

$$\mathrm{d}g\mathrm{d}\xi = \mathrm{d}(T_e\mathsf{L}_ge_1,0)\wedge\cdots\wedge\mathrm{d}(T_e\mathsf{L}_ge_m,0)\wedge\mathrm{d}(0,e_1)\cdots\wedge\mathrm{d}(0,e_m)$$

applied to our $2m$ independent vector fields gives a non-zero constant function, they are identical up to a constant. ∎

C.3. Discussion on the term $\mathrm{ad}_\xi^*\xi$

There are several reasons why the term $\mathrm{ad}_\xi^*\xi$ is required in both optimization dynamics Eq. (3) and sampling dynamics Eq. (4):

1. From the view of mechanics (discussed in Sec. C.2), it comes from the third term of the natural symplectic 2-form Eq. (18).

2. From the view of Riemannian geometry (will be discussed later in Sec. D.1), it is a term from the definition of geodesics.

3. Another technical reason is the term is required to ensure the invariant distribution is correct because the divergence of left-invariant vector fields does not always vanish, see more details in the proof of Thm. 22.

However, despite the necessity of the term $\mathrm{ad}_\xi^*\xi$ on a general left-invariant metric, we can carefully choose the inner product to make it vanish (Lemma 3 and Rmk. 26), and all the above still hold.

C.4. Gibbs as invariant distribution of sampling dynamics

This section will be devoted to showing that the Gibbs distribution $\mathsf{G}\times\mathfrak{g}$ Eq. (2) is an invariant distribution of our sampling dynamics Eq. (4) under mild conditions. Unlike the stronger Assumption 1 we made for convergence, we do not require the special choice of inner product on $\mathfrak{g}$ in Lemma 3 nor the compactness of Lie group in this section.

Definition 20 (Structural constant) *By denoting $\{e_i\}_{i=1}^m$ as a set of orthonormal basis of $\mathfrak{g}$ under inner product $\langle\cdot,\cdot\rangle$, the structural constant of Lie algebra $\mathfrak{g}$ defined as*

$$c_{ij}^k = \langle[e_i,e_j],e_k\rangle$$

Lemma 21 *For any $\xi \in \mathfrak{g}$, we have*

$$\mathrm{div}_g(T_e \mathsf{L}_g \xi) + \mathrm{div}_\xi \left(\mathrm{ad}_\xi^* \xi \right) = 0$$

where $T_e \mathsf{L}_g \xi$ is the left-invariant vector field generated by ξ.

Proof [Proof of Lemma 21] Consider the local chart on G at g given by $\{g \exp(x_i e_i)\}_{i=1}^m$ for $x_i \in [-\varepsilon, \varepsilon]^m$. We have

$$\mathrm{div}_g(T_e \mathsf{L}_g \xi) = \sum_i \partial_i \xi_i + \sum_{ij} \Gamma_{ji}^j \xi_i$$

By Eq. (19), $\Gamma_{ji}^j = 0, \forall i, j$, and the term $\sum_{ij} \Gamma_{ji}^j \xi_i$ vanishes consequently. By our choice of the local coordinate,

$$
\begin{aligned}
\partial_i \xi_i &= \mathcal{L}_{e_i} \langle T_e \mathsf{L}_{\hat{g}} \xi, e_i \rangle |_{\hat{g}=e} \\
&= \langle \mathcal{L}_{e_i} T_e \mathsf{L}_{\hat{g}} \xi, e_i \rangle |_{\hat{g}=e} + \langle T_e \mathsf{L}_{\hat{g}} \xi, \mathcal{L}_{e_i} e_i \rangle |_{\hat{g}=e} \\
&= \langle [e_i, \xi], e_i \rangle + \langle T_e \mathsf{L}_{\hat{g}} \xi, [e_i, e_i] \rangle \\
&= \langle [e_i, \xi], e_i \rangle
\end{aligned}
$$

where $\mathcal{L}$ is the Lie derivaitve. In the first line of equation, $T_e \mathsf{L}_{\hat{g}} \xi$ is the left-invariant vector field generated by ξ. $|_{\hat{g}=e}$ means we evaluate the Lie derivative at e.

After taking summation w.r.t. i, we have $\mathrm{div}_g(T_e \mathsf{L}_g \xi) = \sum_i \partial_i \xi_i = \langle \xi, \sum_k \mathrm{ad}_{e_k}^* e_k \rangle$.

For $\mathrm{div}_\xi \left(\mathrm{ad}_\xi^* \xi \right)$, the divergence is taken in the Euclidean space $\mathfrak{g}$ and a direct calculation gives

$$
\begin{aligned}
\mathrm{div} \, \mathrm{ad}_\xi^* \xi &= \sum_i \partial_i \left(\mathrm{ad}_\xi^* \xi \right)_i \\
&= \sum_{ijk} \partial_i \xi_j \xi_k \langle \mathrm{ad}_{e_j}^* e_k, e_i \rangle \\
&= -\sum_i \langle \xi_j, \mathrm{ad}_{e_i}^* e_i \rangle \\
&= -\mathrm{div}_g(T_e \mathsf{L}_g \xi)
\end{aligned}
$$

∎

Now we are ready to prove Thm. 22. The outline for the proof is: 1) calculate the infinitesimal generator $\mathcal{L}$ by definition; 2) find its adjoint operator $\mathcal{L}^*$ under L^2; 3) verify that the Gibbs distribution is a fixed point of the adjoint of the infinitesimal generator.

Theorem 22 *Suppose the Lie group G is finite dimensional, then the sampling dynamics Eq.(4) has invariant distribution ν_* in Eq. (2).*

Proof [Proof of Thm. 22] We first write down the infinitesimal generator $\mathcal{L}$ for SDE (4). For any $f \in C^2(\mathsf{G} \times \mathfrak{g})$, $\mathcal{L}$ is defined as

$$
\begin{aligned}
\mathcal{L}f(g, \xi) &:= \lim_{\delta \to 0} \frac{\mathbb{E}\left[f(g_\delta, \xi_\delta) | (g_0, \xi_0) = (g, \xi) \right] - f(g, \xi)}{\delta} \\
&= \langle \nabla_g f, T_e \mathsf{L}_g \xi \rangle + \langle \nabla_\xi f, -\gamma \xi + \mathrm{ad}_\xi^* \xi - T_g \mathsf{L}_{g^{-1}}(\nabla U(g)) \rangle + \gamma \Delta_\xi f \\
&= \langle \xi, T_g \mathsf{L}_{g^{-1}} \nabla_g f \rangle - \gamma \langle \nabla_\xi f, \xi \rangle + \langle \nabla_\xi f, \mathrm{ad}_\xi^* \xi \rangle - \langle \nabla_\xi f, T_g \mathsf{L}_{g^{-1}}(\nabla U(g)) \rangle + \gamma \Delta_\xi f
\end{aligned}
$$

We denote the adjoint operator of $\mathcal{L}$ by $\mathcal{L}^*$, i.e., $\mathcal{L}^* : C^2 \to C^2$ satisfying $\int_{\mathsf{G} \times \mathfrak{g}} \nu \mathcal{L} f \mathrm{d}g \mathrm{d}\xi = \int_{\mathsf{G} \times \mathfrak{g}} f \mathcal{L}^* \nu \mathrm{d}g \mathrm{d}\xi$ for any $f, \nu \in C_0^2(\mathsf{G} \times \mathfrak{g})$. By the divergence theorem, we have

$$
\begin{aligned}
&\int_{\mathsf{G} \times \mathfrak{g}} \nu \mathcal{L} f \, \mathrm{d}g \mathrm{d}\xi \\
&= \int_{\mathsf{G} \times \mathfrak{g}} -f \operatorname{div}_g(\nu T_e \mathsf{L}_g \xi) - f \operatorname{div}_\xi \left(\nu \operatorname{ad}_\xi^* \xi \right) + f \operatorname{div}_\xi \left(\nu T_g \mathsf{L}_{g^{-1}}(\nabla U(g)) \right) \mathrm{d}g \mathrm{d}\xi \\
&\quad + \gamma \int_{\mathsf{G} \times \mathfrak{g}} f \operatorname{div}_\xi(\nu \xi) + f \Delta_\xi \nu \mathrm{d}g \mathrm{d}\xi
\end{aligned}
$$

Here $T_e \mathsf{L}_g \xi$ stands for the left-invariant vector filed on G generated by ξ. As a result, we have

$$
\begin{aligned}
\mathcal{L}^* \nu &= -\operatorname{div}_g(\nu T_e \mathsf{L}_g \xi) - \operatorname{div}_\xi \left(\nu \operatorname{ad}_\xi^* \xi \right) + \operatorname{div}_\xi \left(\nu T_g \mathsf{L}_{g^{-1}}(\nabla U(g)) \right) + \gamma \left(\operatorname{div}_\xi(\nu \xi) + \Delta_\xi \nu \right) \\
&= -\langle \nabla_g \nu, T_e \mathsf{L}_g \xi \rangle - \langle \nabla_\xi \nu, \operatorname{ad}_\xi^* \xi \rangle + \langle \nabla_\xi \nu, T_g \mathsf{L}_{g^{-1}}(\nabla U(g)) \rangle + \gamma \operatorname{div}_\xi(\nu \xi + \nabla_\xi \nu) \\
&\quad - \nu \operatorname{div}_g(T_e \mathsf{L}_g \xi) - \nu \operatorname{div}_\xi \left(\operatorname{ad}_\xi^* \xi \right)
\end{aligned}
$$

We emphasize that the divergence of the left-invariant vector does not necessarily vanish, and Lemma 21 shows $\operatorname{div}_g(T_e \mathsf{L}_g \xi) + \operatorname{div}_\xi \left(\operatorname{ad}_\xi^* \xi \right) = 0$, which means the term $\operatorname{ad}_\xi^* \xi$ is necessary to cancel with the divergence of left-invariant vector field and ensure the invariant distribution is correct.

$$
\mathcal{L}^* \nu = -\langle \nabla_g \nu, T_e \mathsf{L}_g \xi \rangle - \langle \nabla_\xi \nu, \operatorname{ad}_\xi^* \xi \rangle + \langle \nabla_\xi \nu, T_g \mathsf{L}_{g^{-1}} \nabla U(g) \rangle + \gamma \operatorname{div}_\xi(\nu \xi + \nabla_\xi \nu)
$$

The last step is verifying that ν_* given in Eq. (2) is a fixed point of $\mathcal{L}^*$. By direct calculation, we have $\nabla_g \nu_* = \nu_* \nabla U$ and $\nabla_\xi \nu_* = -\nu_* \xi$. Together with expression $\mathcal{L}^*$, we have $\mathcal{L}^* \nu_* = 0$, which means ν_* given in Eq. (2) is an invariant distribution. ∎

The following remark gives the left Haar measure an intuition.

Remark 23 (Left Haar measure) *The base measure $\mathrm{d}g$ we used is called the 'left Haar measure'. Roughly speaking, if we have a 'measure' at e (rigorously speaking, a volume form at e), we can expand it to the whole Lie group by left multiplication, thanks to the group structure. This measure is the left Haar measure, rigorously defined as the measure that is invariant under the pushforward by left multiplication, which is unique up to a constant scaling factor. The reason why it is 'left' is because our SDE Eq. (1) and later (4) depends on left trivialization and our metric on G is left-invariant.*

Appendix D. Riemannian structure on Lie groups with left-invariant metric

D.1. More discussion on Lemma 3

In the beginning of Sec. 2.3, we mentioned that Lemma 3 means the Riemannian structure and the group structure are 'compatible'. Here we provide more details: it means the exponential map from the group structure and the exponential map from the Riemannian structure are the same.

Definition 24 (Exponential map (group structure)) *The exponential map* $\exp : \mathfrak{g} \to$ G *is given by* $\exp(\xi) = \gamma(1)$ *where* $\gamma : \mathbb{R} \to \mathsf{G}$ *is the unique one-parameter subgroup of* G *whose tangent vector at the group identity e is equal to ξ.*

Definition 25 (Exponential map (Riemannian structure)) *The exponential map* $\exp :$ $T_g \mathsf{G} \to \mathsf{G}$ *is given by* $\exp_g(T_e \mathsf{L}_g \xi) = \gamma(1)$ *where* $\gamma : \mathbb{R} \to \mathsf{G}$ *is the unique geodesic satisfying* $\gamma(0) = g$ *with initial tangent vector* $\gamma'(0) = T_e \mathsf{L}_g \xi$.

To compare the two exponential maps, we can write down the ODEs characterising these two exponential maps. Starting from g_0 with initial direction $T_e \mathsf{L}_{g_0} \xi_0$, the trajectory of two exponential maps $\exp_{g_0}(t\xi_0), t \in [0, 1]$ are given by the ODEs with initial condition $g(0) = g_0$ and $\xi(0) = \xi_0$:

The exponential map from the group structure (Def. 24) is given by

$$\begin{cases} \dot{g} = g\xi \\ \dot{\xi} = 0 \end{cases}$$

The exponential map from the Riemannian structure (Def. 25) is given by

$$\begin{cases} \dot{g} = g\xi \\ \dot{\xi} = \mathrm{ad}_\xi^* \xi \end{cases}$$

If we choose the inner product that makes ad skew-adjoint (Lemma 3), the two exponential maps are identical by comparing their ODEs. This means the Riemannian structure and the group structure are compatible and will be our default choice in the following. As a result, we will no longer differentiate the two exponential.

Remark 26 (Explicit expression of the inner product in Lemma 3) *The condition that* ad *is skew adjoint (Lemma 3) is equal to the requirement that the metric is bi-invariant, i.e., a metric that is both left-invariant and right-invariant. A left-invariant metric is not always bi-invariant because of the non-commutativity of the group structure. However, a connected Lie group admits such a bi-invariant metric if and only if it is isomorphic to the Cartesian product of a compact group and a commutative group (Milnor, 1976). On a compact Lie group, an explicit expression for a bi-invariant metric is*

$$\langle u, v \rangle = \int_{\mathsf{G}} \left(\mathrm{Ad}_g(u), \mathrm{Ad}_g(v) \right) \vartheta(\mathrm{d}g)$$

where $(\cdot, \cdot)$ is an arbitrary inner product on $\mathfrak{g}$ and ϑ is the **right** *Haar measure. See Milnor (1976); Lezcano-Casado and Martinez-Rubio (2019) for more details.*

D.2. Connection and curvature of Lie groups with left-invariant metric

Given left-invariant vector fields X and Y, the Levi-Civita connection on Lie groups with a left-invariant metric is given by (Milnor, 1976):

$$\nabla_X Y = \frac{1}{2} \left([X, Y] - \mathrm{ad}_X^* Y - \mathrm{ad}_Y^* X \right)$$

which leads to the following Christoffel symbols

$$\Gamma_{ij}^k := \langle \nabla_{e_i} e_j, e_k \rangle = \frac{1}{2} \left(c_{ij}^k - c_{ik}^j - c_{jk}^i \right) \tag{19}$$

Here e_i stands for the left-invariant vector field generated by e_i.

Under the inner product in Lemma 3, Levi-Civita connection has a simpler expression given by $\nabla_X Y = \frac{1}{2}[X, Y]$, i.e., Levi-Civita connection is half of the Lie derivative. The Christoffel symbols becomes $\Gamma_{ij}^k = \frac{1}{2} c_{ij}^k$.

The Riemannian curvature tensor is defined by $R(X, Y)Z := \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z$. The sectional curvature on a general manifold is defined as $\kappa(X, Y) = \frac{\langle R(X,Y)Y,X \rangle}{\langle X,X \rangle \langle Y,Y \rangle - \langle X,Y \rangle^2}$

Under the condition that the operator ad^* is skew-adjoint (Lemma 3), and X, Y are orthonormal, the simplified Levi-Civita expression $\nabla_X Y = \frac{1}{2}[X, Y]$ gives us

$$\kappa(e_i, e_j) = \frac{1}{4} \|[e_i, e_j]\|^2 \tag{20}$$

Therefore the sectional curvature is non-negative. If we denote K as the upper bound for the absolute value of sectional curvature, i.e., $K = \sup |\kappa|$, and let

$$C := \max_{\|X\|_2 = 1} \|\mathrm{ad}_X\|_{op} \tag{21}$$

Then Eq. (20) leads to $K = \frac{1}{4}C^2$. Note $C = 0$ in the flat Euclidean space.

D.3. About L-smoothness on Lie groups

The commonly used geodesic L-smooth on a manifold $\mathcal{M}$ is given by the following definition (e.g., Zhang and Sra, 2016, Def. 5):

Definition 27 (Geodesically L-smooth) $U : \mathsf{G} \to \mathbb{R}$ *is geodesically L-smooth if for any* $g, \hat{g} \in \mathcal{M}$,

$$\left\| \nabla U(g) - \Gamma_{\hat{g}}^g \nabla U(\hat{g}) \right\| \le L d(g, \hat{g})$$

where $\Gamma_{\hat{g}}^g$ is the parallel transport from $\hat{g}$ to g.

The following lemma, quoted from Kong and Tao (2024) just for completeness, shows it is equal to the commonly used geodesic L-smoothness under the left-invariant metric in Lemma 3:

Lemma 28 (Equivalence of trivialized smoothness and geodesic smoothness) *Under the assumption of ad_X^* is skew-adjoint $\forall X \in \mathfrak{g}$, Assumption 2 is identical to Def. 27.*

Proof [Proof of Lemma 28, from Kong and Tao (2024)] For any $g, \hat{g} \in \mathsf{G}$, consider the shortest geodesic $\phi : [0, 1] \to \mathsf{G}$ connecting g and $\hat{g}$ and denote $\xi = T_g \mathsf{L}_{g^{-1}} \nabla U(g)$. Using the condition ad is skew-adjoint, we have $\ddot{\phi}(t) = 0$ and $T_e \mathsf{L}_{\phi(t)} \xi$ is parallel along ϕ by checking the condition for parallel transport (Guigui and Pennec, 2021, Thm. 1):

$$\frac{\mathrm{d}}{\mathrm{d}t} \xi = 0 = -\frac{1}{2} \left[T_{\phi(t)^{-1}} \mathsf{L}_{\phi(t)^{-1-1}} \dot{\phi}(t), \xi \right]$$

This tells that

$$T_g\mathsf{L}_{g^{-1}}\Gamma_g^{\hat{g}}\nabla f(g) = T_{\hat{g}}\mathsf{L}_{\hat{g}^{-1}}\nabla f(\hat{g})$$

Together with the metric is left-invariant, we have

$$\left\|T_g\mathsf{L}_{g^{-1}}\nabla U(g) - T_{\hat{g}}\mathsf{L}_{\hat{g}^{-1}}\nabla U(\hat{g})\right\| \le Ld(g,\hat{g})$$

which is identical to Assumption 2. ∎

D.4. Non-commutativity of Lie groups

Comparing with the Euclidean space, Lie groups lack of commutativity, i.e., for $g,\hat{g} \in \mathsf{G}$, $g\hat{g}$ and $\hat{g}g$ are not necessarily equal. This can also be characterized by the non-trivial Lie bracket. This non-commutativity leads to the fact that $\exp(X)\exp(Y) \ne \exp(X + Y)$. An explicit expression for $\log(\exp(X)\exp(Y))$ is given by Dynkin's formula (Dynkin, 2000). In the matrix group case, exp is matrix exponentiation and multiplication is matrix multiplication, and for matrices A, B, we no longer have $\exp(A)\exp(B) = \exp(A + B)$.

We will not use Dynkin's formula directly but only 2 corollaries of it, Cor. 29 and 30. By defining power series

$$p(x) := \frac{x}{1 - \exp(-x)} = \sum_{k=0}^{\infty} \frac{B_k(1)}{k!}x^k \tag{22}$$

and its inverse

$$p^{-1}(x) := \frac{1 - \exp(-x)}{x} = \sum_{k=0}^{\infty} \frac{(-1)^k}{(k + 1)!}x^k \tag{23}$$

where B_n are Bernoulli polynomials, we have

Corollary 29 (Differential of exponential) *The differential of matrix exponential is given by*

$$\mathrm{d}\exp_X(Y) = T_e\mathsf{L}_{\exp X}\left[p^{-1}(\mathrm{ad}_X)Y\right], \quad \forall X, Y \in \mathfrak{g} \tag{24}$$

Thm. 2.2 in Bröcker and Tom Dieck (2013) shows that the group exponential is surjective on compact connected Lie groups. Thus, We can define **logarithm** $\log_g : \mathsf{G} \to T_g\mathsf{G}$ as the inverse of an exponential map. Since the exponential map is not injective in general cases, its inverse, logarithm is not uniquely defined. In this paper, we use all the log to denote the one corresponding to the minimum geodesic, i.e., $d(g, h) = \left\|\log_g h\right\|, \forall g, h$. In the cases that the minimum geodesic is not unique, we use log to denote any one of them. We omit the subscript when the base point is identity, i.e., $\log := \log_e$. Note that we do not require the uniqueness of the geodesic connecting 2 points.

We also need to consider the differential of the logarithm. However, since the exponential map is not an injection, the logarithm is not always differentiable. We denote the set that the differential of logarithm does not exist as

$$N := \left\{g \in \mathsf{G} : \mathrm{d}\log_g \text{ is not well defined}\right\} \tag{25}$$

For $g \in \mathsf{G}\backslash N$, we have the following expression for $\mathrm{d}\log_g$

Corollary 30 (Differential of logarithm) *For any $g \notin N$, any $\xi \in \mathfrak{g}$, we have the differential of $\log$ is given by $\operatorname{d}\log_g : T_g \mathsf{G} \to \mathfrak{g}$ is given by*

$$\operatorname{d}\log_g(T_e \mathsf{L}_g \xi) = p(\operatorname{ad}_{\log_g})\xi$$

The main property of $\operatorname{d}\log$ we are going to use is the following corollary, which is useful later in the proof of Lemma 6.

Corollary 31 *Under the choice of inner product in Lemma 3, for any $g \notin N$ and $\xi \in \mathfrak{g}$, we have*

$$\left\langle \operatorname{d}\log_g(T_e \mathsf{L}_g \xi), \log g \right\rangle = \langle \log g, \xi \rangle \tag{26}$$

$$\left\langle \operatorname{d}\log_g(T_e \mathsf{L}_g \xi), \xi \right\rangle \leq \|\xi\|^2 \tag{27}$$

Proof [Proof of Cor. 31] Proof for Eq. (26): By Cor. 30, we have

$$\left\langle \operatorname{d}\log_g(T_e \mathsf{L}_g \xi), \log g \right\rangle = \left\langle p(\operatorname{ad}_{\log_g})\xi, \log g \right\rangle$$

Using the fact that ad^* is skew-adjoint, $\left\langle \left(\operatorname{ad}_{\log_g}\right)^k \xi, \log_g \right\rangle = \left\langle \left(\operatorname{ad}_{\log_g}\right)^{k-1} \xi, \operatorname{ad}^*_{\log_g} \log_g \right\rangle = 0$
for any $k \geq 1$. Since the 0-order term in p is 1,

$$\left\langle \operatorname{d}\log_g(T_e \mathsf{L}_g \xi), \log g \right\rangle = \langle \xi, \log g \rangle$$

Proof for Eq. (27): By Cor. 30, we have

$$\left\langle \operatorname{d}\log_g(T_e \mathsf{L}_g \xi), \xi \right\rangle = \left\langle p(\operatorname{ad}_{\log_g})\xi, \xi \right\rangle$$

$\operatorname{ad}_{\log_g(g^*)}$ is a linear map from $\mathfrak{g}$ to itself. Since the operator is skew-adjoint, it has all eigenvalues pure imagine or 0. By the assumption that $p(\operatorname{ad}_{\log_g(g^*)})$ converges, we have all the eigenvalues in the interval $(-2\pi i, -2\pi i)$. Since $\Re[p(xi)] = \frac{x\sin(x)}{2(1-\cos(x))} \leq 1$ for $x \in (-2\pi, 2\pi)$, the eigenvalues of $p(\operatorname{ad}_{\log_g(g^*)})$ are smaller or equal to 1, which gives us Eq. (27). ∎

By Cor. 30, the set N in Eq. (25) is also equal to the set where the power series $p(\operatorname{ad}_{\log_g})$ does not converge. We have the following lemma showing this set is zero-measured under the left Haar measure.

Lemma 32 *By choosing the inner product in Lemma 3, the pushforward measure of $\operatorname{d}\xi$ on $\mathfrak{g}$ by $\exp$ is absolute continuous w.r.t. $\operatorname{d}g$, i.e.,*

$$\operatorname{d}g \ll \exp_\sharp(\operatorname{d}\xi)$$

Proof The pushforward measure by $\exp : \mathfrak{g} \to \mathsf{G}$ is given by $\operatorname{d}g = \det(p^{-1}(\operatorname{ad}_X))\operatorname{d}\xi$ where p^{-1} is defined in Eq. 23. Since we have $\|p^{-1}(xi)\| \leq 1, \forall x \in \mathbb{R}$, we have $\det(p^{-1}(\operatorname{ad}_\xi))$ is upper bounded, which means $\exp$ pushes any zero-measured set in $\mathfrak{g}$ to a zero-measured set in G, i.e., $\operatorname{d}g \ll \exp_\sharp(\operatorname{d}\xi)$. ∎

Lemma 33 $p(\mathrm{ad}_{\log g})$ *is well defined for* $g \in \mathsf{G}$ *almost everywhere under measure* $\mathrm{d}g$.

Proof [Proof of Lemma 33] By the expression of p in Eq. (22), we have $p(x)$ diverges when $x \neq 0$ and $\exp(-x) = 1$, i.e., $x = 2k\pi i$ for $k \neq 0$. As a result, $p(\mathrm{ad}_\xi)$ is well defined if $\xi \in P$,

$$P := \left\{ \xi \in \mathfrak{g} : \lambda \neq 2k\pi i, \forall \lambda \in \sigma(\mathrm{ad}_\xi), \forall k \neq 0 \right\}$$

where $\sigma(\mathrm{ad}_\xi)$ denotes the set of all eigenvalues of the linear operator ad_ξ.

The Lebesgue measure of P is given by

$$\int_{\mathfrak{g}} \mathbb{1}_P(\xi)\mathrm{d}\xi$$
$$= \mathrm{Area}(\partial B_1(0)) \int_{\partial B_1(0)} \int_0^\infty \mathbb{1}_P(te)t^{d-1}\mathrm{d}t\mathrm{d}e$$

Since we have the operator ad_ξ is linear in ξ, which gives us that $\sigma(\mathrm{ad}_{t\xi}) = t\sigma(\mathrm{ad}_\xi)$ (σ is the set of eigenvalues for operators). Using the assumption that $\mathfrak{g}$ is finite dimensional, we have for a fixed $e \in \mathfrak{g}$, only countable many t makes $te \in P$, which means $\int_0^\infty \mathbb{1}_P(te)t^{d-1}\mathrm{d}t = 0$, this gives P has zero measure. By Lemma 32, $\exp(P)$ has measure 0 under left Haar measure on G. ∎

This lemma proves that we have d log well defined almost everywhere on G. Later in our proof of convergence using coupling method, the lemma ensures that we can ignore the pairs $(g, \hat{g})$ that $\mathrm{d}\log \hat{g}^{-1}g$ do not exist when we consider the coupling of 2 trajectories in $\mathsf{G} \times \mathfrak{g}$ since a zero-measured set will be negligible when taking expectation. See the proof of Thm. 7.

Appendix E. Absolute continuity w.r.t. $\mathrm{d}g\mathrm{d}\xi$

As shown in Lemma 33, the set that d log is not well-defined is zero-measured under the left Haar measure. To rule out this pathological part, we need to show that the density for sampling SDE Eq. (4) (ν_t) and the numerical scheme Eq. (14) ($\tilde{\nu}_k$) are both absolute continuous w.r.t. $\mathrm{d}g\mathrm{d}\xi$, and then N is also zero-measured under ν_t and $\tilde{\nu}_k$.

The KL divergence between two probability measures ν_1 and ν_2 is defined by

$$\mathrm{KL}\left(\nu_1 \| \nu_2\right) := \int_{\mathsf{G} \times \mathfrak{g}} \ln\left(\frac{\nu_1}{\nu_2}\right) d\nu_1$$

A direct conclusion is that ν_1 is absolute continuous w.r.t. ν_2 if $\mathrm{KL}\left(\nu_1\|\nu_2\right) < \infty$. Using the data processing inequality, we are going to show $\mathrm{KL}\left(\nu_t\|\nu_*\right) < \infty$ and $\mathrm{KL}\left(\tilde{\nu}_k\|\nu_*\right) < \infty$, which further induces ν_t and $\tilde{\nu}_k$ is absolute continuous w.r.t. ν_* (and also $\mathrm{d}g\mathrm{d}\xi$), respectively.

Lemma 34 (Data processing inequality) *For any Markov transitional kernel P, and any probability distributions ν_1 and ν_2,*

$$\mathrm{KL}\left(\nu_1 P \| \nu_2 P\right) \leq \mathrm{KL}\left(\nu_1 \| \nu_2\right)$$

Corollary 35 *Given initial condition (g_0, ξ_0) is absolute continuous w.r.t. $\mathrm{d}g\mathrm{d}\xi$, (g_t, ξ_t) (evolution of SDE Eq. (4)) and $(\tilde{g}_k, \tilde{\xi}_k)$ (evolution of numerical scheme Eq. (14) with step size h) are both absolute continuous w.r.t. $\mathrm{d}g\mathrm{d}\xi$ for any $t \geq 0$.*

Although Cor. 35 requires the absolute continuity of (g_0, ξ_0) to ensure the density $(\tilde{g}_k, \tilde{\xi}_k)$ is absolutely continuous in later steps, this can be challenging in practice and we do not require our initialization to be absolute continuous w.r.t. $\mathrm{d}g\mathrm{d}\xi$ in our proposed Algo. 1 for the following reason:

Remark 36 (About absolute continuity of initialization) *In our implementation of Algo. 1, we use the simplest initialization that g_0 is an arbitrary point and $\xi_0 = 0$ but did not require the absolute continuouity as in the proof of convergence rate in Thm. 16. This is because the absolute continuity will be automatic after several iterations: In the first iteration, we first update ξ_0 by Eq. (12), which is an Ornstein–Uhlenbeck process and at time h, the density of $\tilde{\xi}_1$ is absolute continuous w.r.t. $\mathrm{d}\xi$. As a result, since $g_1 = g_0 \exp(h\xi_1)$, we have the density of $\tilde{g}_1$ is densities absolute continuous w.r.t. $\mathrm{d}g$ (Lemma 32). In the second iteration, since noise is added to $\tilde{\xi}_1$ again, we have the joint distribution (g_1, ξ_2) is already absolute continuous w.r.t. $\mathrm{d}g\mathrm{d}\xi$. And by the data processing inequality (Lemma 34), we have from then on, the density will always be absolute continuous w.r.t. $\mathrm{d}g\mathrm{d}\xi$. As a result, the requirement for initialization does not matter in practice and will be satisfied in two iterations automatically.*

Appendix F. Proof of Thm. 7

F.1. Semi-martingale decomposition

In this section, we use the following shorthand notation

$$
\begin{aligned}
r_t &:= r((g_t, \xi_t), (\hat{g}_t, \hat{\xi}_t)) \\
\rho_t &:= \rho((g_t, \xi_t), (\hat{g}_t, \hat{\xi}_t)) \\
G_t &:= G(\xi_t, \hat{\xi}_t) \\
\delta_t &:= \delta((g_t, \xi_t), (\hat{g}_t, \hat{\xi}_t)) \\
K_t &:= K((g_t, \xi_t), (\hat{g}_t, \hat{\xi}_t))
\end{aligned}
$$

where (g_t, ξ_t) and $(\hat{g}_t, \hat{\xi}_t)$ are coupled using reflection coupling in Sec. 4.1.

The function $K : (\mathsf{G} \times \mathfrak{g})^2 \to \mathbb{R}$ is defined by

$$
\begin{aligned}
K &:= \Big[cf(r) + \Big(\alpha\gamma\|Q\| - \theta L\gamma^{-1}\|Z\| + \mathbb{1}_{\{\hat{g}_t^{-1}g_t \in N\}} \delta \Big) f'_-(r) + 4\gamma^{-1}\mathrm{rc}(Z, \mu)^2 f''(r) \Big] \big(1 + \beta\|\mu\|^2\big) \\
&\quad + \beta f(r) \big[-2\gamma\|\mu\|^2 + 2L\|\mu\|\|Z\| + 8\gamma\,\mathrm{rc}(Z, \mu)^2\|\mu\|^2 \big] \\
&\quad + 16\beta\|\mu\|\,\mathrm{rc}(Z, \mu)^2 f'_-(r)
\end{aligned}
$$

where $\delta : (\mathsf{G} \times \mathfrak{g})^2 \to \mathbb{R}$ is defined by $\delta := (\alpha + 2)\|\mu\| + \alpha\gamma\|Z\| - \alpha\gamma\|Q\|$ and $\theta \in \mathbb{R}$ is defined by $\theta := \alpha\gamma^2 L^{-1} - 1$. This definition for K comes from the calculation of semi-martingale decomposition in Lemma 6.

Proof [Proof of Lemma 6] **Bound the time derivative for $\|Z_t\|$:** For $\hat{g}_t^{-1}g_t \notin N$, the paths of the process (Z_t) are almost surely continuously differentiable with derivative

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} Z_t &= \mathrm{d}\log_{\hat{g}_t^{-1}g_t} \frac{\mathrm{d}}{\mathrm{d}t} g - \mathrm{d}\log_{\hat{g}_t^{-1}g_t} \frac{\mathrm{d}}{\mathrm{d}t}\hat{g} \\
&= \mathrm{d}\log_{\hat{g}_t^{-1}g_t} (\xi_t - \hat{\xi}_t)
\end{aligned}
$$

Using the property of d log in Cor. 31,

$$
\frac{\mathrm{d}}{\mathrm{d}t}\|Z_t\| = \frac{1}{\|Z_t\|}\left\langle Z_t, \frac{\mathrm{d}}{\mathrm{d}t}Z_t\right\rangle = \frac{1}{\|Z_t\|}\left\langle Z_t, \mathrm{d}\log_{\hat{g}^{-1}g}(\xi_t - \hat{\xi}_t)\right\rangle
$$

$$
= \frac{1}{\|Z_t\|}\langle Z_t, \mu_t\rangle = \frac{1}{\|Z_t\|}\langle Z_t, -\gamma Z_t + \gamma Q_t\rangle
$$

$$
\leq -\gamma\|Z_t\| + \gamma\|Q_t\|
$$

When $\hat{g}_t^{-1}g_t \in N$, we won't have $\frac{\mathrm{d}}{\mathrm{d}t}Z_t$. However, a bound for $\frac{\mathrm{d}}{\mathrm{d}t}\|Z_t\|$ can still be derived by the triangle inequality:

$$
\frac{\mathrm{d}}{\mathrm{d}t}\|Z_t\| \leq \|\mu_t\|
$$

In summary,

$$
\frac{\mathrm{d}}{\mathrm{d}t}\|Z_t\| \leq -\gamma\|Z_t\| + \gamma\|Q_t\| + \mathbb{1}_{\{\hat{g}_t^{-1}g_t\in N\}}\left(\|\mu_t\| + \gamma\|Z_t\| - \gamma\|Q_t\|\right)
$$

Semimartingale decomposition for $\|Q_t\|$: Suppose $\|Q_t\|$ satisfies the following inequality almost surely

$$
\|Q_t\| \leq \|Q_0\| + A_t^Q + \tilde{M}_t^Q \qquad \text{for all } t \geq 0,
$$

where (A_t^Q) and $(\tilde{M}_t^Q)$ are the absolutely continuous process and the martingale to be determined.

By our choice of coupling in Sec. 4.1, μ_t follows the SDE

$$
\mathrm{d}\mu_t = -\gamma\mu_t\mathrm{d}t - \left(T_{g_t}\mathsf{L}_{g_t^{-1}}\nabla U(g_t) - T_{\hat{g}_t}\mathsf{L}_{\hat{g}_t^{-1}}\nabla U(\hat{g}_t)\right)\mathrm{d}t + \sqrt{8\gamma}\,\mathrm{rc}(Z_t, \mu_t)e_t e_t^\top \mathrm{d}W_t^{\mathrm{rc}} \tag{28}
$$

When $\hat{g}_t^{-1}g_t \in N$, by $\frac{\mathrm{d}}{\mathrm{d}t}\|Z_t\| \leq \|\mu_t\|\mathrm{d}t$, we have

$$
\frac{\mathrm{d}}{\mathrm{d}t}A_t^Q \leq 2\|\mu_t\|\mathrm{d}t + \gamma^{-1}\left\|T_{g_t}\mathsf{L}_{g_t^{-1}}\nabla U(g_t) - T_{\hat{g}_t}\mathsf{L}_{\hat{g}_t^{-1}}\nabla U(\hat{g}_t)\right\|
$$

When $\hat{g}_t^{-1}g_t \notin N$, by the fact that $Q = Z + \gamma^{-1}\mu$, The process (Q_t) satisfies the following SDE:

$$
\mathrm{d}Q_t = \mathrm{d}Z_t - \mu_t\mathrm{d}t - \gamma^{-1}\left(T_{g_t}\mathsf{L}_{g_t^{-1}}\nabla U(g_t) - T_{\hat{g}_t}\mathsf{L}_{\hat{g}_t^{-1}}\nabla U(\hat{g}_t)\right)\mathrm{d}t + \sqrt{8\gamma^{-1}}\,\mathrm{rc}(Z_t, \mu_t)e_t e_t^\top \mathrm{d}W_t^{\mathrm{rc}}.
$$

Notice that by (6), the noise coefficient vanishes if $Q_t = 0$. Cor. 31 gives us

$$
\left\langle\frac{\mathrm{d}}{\mathrm{d}t}Z_t - \mu_t, Q_t\right\rangle = \left\langle\mathrm{d}\log_{\hat{g}_t^{-1}g_t}\mu_t - \mu_t, \log_{\hat{g}_t^{-1}g_t} + \gamma^{-1}\mu_t\right\rangle
$$

$$
= \left\langle\mathrm{d}\log_{\hat{g}_t^{-1}g_t}\mu_t, \log_{\hat{g}_t^{-1}g_t}\right\rangle + \left\langle\mathrm{d}\log_{\hat{g}_t^{-1}g_t}\mu_t, \gamma^{-1}\mu_t\right\rangle - \left\langle\mu_t, \log_{\hat{g}_t^{-1}g_t}\right\rangle - \left\langle\mu_t, \gamma^{-1}\mu_t\right\rangle
$$

$$
\leq 0
$$

and we have in this case

$$
\frac{\mathrm{d}}{\mathrm{d}t}A_t^Q \leq \gamma^{-1}\left\|T_{g_s}\mathsf{L}_{g_s^{-1}}\nabla U(g_s) - T_{\hat{g}_s}\mathsf{L}_{\hat{g}_s^{-1}}\nabla U(\hat{g}_s)\right\|
$$

35

Therefore, combining the 2 cases and apply Itô's formula to get

$$A_t^Q = \gamma^{-1} \int_0^t \left\| T_{g_s} \mathsf{L}_{g_s^{-1}} \nabla U(g_s) - T_{\hat{g}_s} \mathsf{L}_{\hat{g}_s^{-1}} \nabla U(\hat{g}_s) \right\| + 2\mathbb{1}_{\{\hat{g}_s^{-1} g_s \in N\}} \|\mu_s\| \mathrm{d}s$$

$$\tilde{M}_t^Q = \sqrt{8\gamma^{-1}} \int_0^t \mathrm{rc}(Z_s, \mu_s) e_s^\top \mathrm{d}W_s^{\mathrm{rc}}$$

Notice that there is no Itô correction, because $\partial_{q/\|q\|}^2 \|q\| = 0$ for $q \neq 0$ in both cases and the noise coefficient vanishes for $Q_t = 0$.

By the L-smoothness of U, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} A_t^Q \leq L\gamma^{-1} \|Z_t\| + 2\mathbb{1}_{\{\hat{g}_s^{-1} g_s \in N\}} \|\mu_s\|$$

Semimartingale decomposition for $f(r_t)$: By the definition of r_t in Eq. (8)

$$r_t := \alpha \|Z_t\| + \|Q_t\| = \|Q_0\| + \alpha \|Z_t\| + A_t^Q + \tilde{M}_t^Q$$

where $t \mapsto \alpha \|Z_t\| + A_t^Q$ is almost surely absolutely continuous with time derivative upper bounded by

$$\frac{\mathrm{d}}{\mathrm{d}t} (\alpha \|Z_t\| + A_t^Q) \leq -(\alpha\gamma - L\gamma^{-1}) \|Z_t\| + \alpha\gamma \|Q_t\| + \mathbb{1}_{\{\hat{g}^{-1} g \in N\}} \delta_t$$

$$= -\theta L\gamma^{-1} \|Z_t\| + \alpha\gamma \|Q_t\| + \mathbb{1}_{\{\hat{g}_t^{-1} g_t \in N\}} \delta_t$$

Since by assumption, f is concave and C^2, we can now apply the Ito-Tanaka formula to $f(r_t)$ and obtain a semi-martingale decomposition

$$e^{ct} f(r_t) \leq f(r_0) + \tilde{A}_t^f + \tilde{M}_t^f$$

with the absolute continuous process $(\tilde{A}_t)$ and the martingale $\tilde{M}_t^f$ given by

$$\mathrm{d}\tilde{A}_t^f = e^{ct} \left(cf(r_t) + \left(\alpha\gamma \|Q_t\| - \theta L\gamma^{-1} \|Z_t\| + \mathbb{1}_{\{\hat{g}_t^{-1} g_t \in N\}} \delta_t \right) f_-'(r_t) + 4\gamma^{-1} \mathrm{rc}(Z_t, \mu_t)^2 f''(r_t) \right) dt$$

$$\tilde{M}_t^f = \sqrt{8\gamma^{-1}} \int_0^t e^{cs} f_-'(r_s) \mathrm{rc}(Z_s, W_s) e_s^\top dW_s^{\mathrm{rc}}$$

Semimartingale decomposition for G_t: By the SDE for μ_t in Eq. (28), the SDE for μ_t gives the decomposition for G_t:

$$G_t = G_0 + A_t^G + \tilde{M}_t^G \qquad \text{for all } t \geq 0,$$

where (A_t^G) and $(\tilde{M}_t^G)$ are the absolutely continuous process and the martingale given by

$$A_t^\mu = \int_0^t -2\gamma \|\mu_s\|^2 - 2\mu_s^\top \cdot \left(T_{g_s} \mathsf{L}_{g_s^{-1}} \nabla U(g_s) - T_{\hat{g}_s} \mathsf{L}_{\hat{g}_s^{-1}} \nabla U(\hat{g}_s) \right) + 8\gamma \langle \mu_s, e_s \rangle^2 \mathrm{rc}(Z_s, \mu_s)^2 ds,$$

$$\tilde{M}_t^\mu = \sqrt{32\gamma} \int_0^t \langle \mu_s, e_s \rangle \mathrm{rc}(Z_s, \mu_s) e_s^\top dW_s^{\mathrm{rc}}.$$

Using the L-smoothness of U, we have

$$A_t^\mu \leq -2\gamma \|\mu_t\|^2 + 2L \|Z_t\| \|\mu_t\| + 8\gamma \mathrm{rc}(Z_t, \mu_t)^2 \|\mu_t\|^2$$

Semimartingale decomposition for $e^{ct}\rho_t$: We combine the semimartingale decomposition for $e^{ct}f(r_t)$ and $\|\mu_t\|$ to have the following semimartingale decomposition for $e^{ct}\rho_t$

$$e^{ct}\rho_t = e^{ct}f(r_t)\left(1 + \beta\|\mu_t\|^2\right) = \rho_0 + M_t + A_t$$

where the continuous process A_t satisfies $\frac{\mathrm{d}}{\mathrm{d}t}A_t \le e^{ct}K_t$.

$$
\begin{aligned}
K_t &= \left[cf(r_t) + \left(\alpha\gamma\|Q_t\| - \theta L\gamma^{-1}\|Z_t\| + \mathbb{1}_{\{\hat{g}_t^{-1}g_t \in N\}}\delta_t\right)f'_-(r_t) + 4\gamma^{-1}\operatorname{rc}(Z_t,\mu_t)^2 f''(r_t)\right]\left(1 + \beta\|\mu_t\|^2\right) \\
&\quad + \beta f(r_t)\left[-2\gamma\|\mu_t\|^2 + 2L\|\mu_t\|\|Z_t\| + 8\gamma\operatorname{rc}(Z_t,\mu_t)^2\|\mu_t\|^2\right] \\
&\quad + 16\beta\|\mu_t\|\operatorname{rc}(Z_t,\mu_t)^2 f'_-(r_t)
\end{aligned}
$$

This gives us the definition for $K : (\mathsf{G}\times\mathfrak{g})^2 \to \mathbb{R}$ at the beginning of Sec. F.1 ∎

F.2. Conditions for contractivity

In the following lemma, we find the sufficient condition for f, η (α), β, R to ensure $K_t \le 0$. Note that we ignore the part $\mathbb{1}_{\{\hat{g}^{-1}g \in N\}}\delta_t$ because we has shown it has zero measure at any time (Cor. 35) and will not affect the contraction under W_ρ in Thm. 7 later.

Lemma 37

$$K_t \le \epsilon\alpha\gamma\left(1 + \beta\|\mu_t\|^2\right)$$

when $\hat{g}^{-1}g \notin N$ if the following conditions Eq. (29), (32), (31) are satisfied:

$$A_0 f(r) + A_1 r f'_-(r) + A_2 f''(r) \le 0, \quad \forall r \in [0,R] \tag{29}$$

for

$$
\begin{cases}
A_0 := c + \beta\left[6\gamma^3 R^2 + 2\gamma LRD\right] \\
A_1 := \alpha\gamma + 16\beta\gamma\max\{1,\alpha^{-1}\} \\
A_2 := 4\gamma^{-1}
\end{cases}
\tag{30}
$$

$$c - \beta(2\gamma - c - 2LD\gamma^{-1}R^{-1})\gamma^2 R^2 \le 0 \tag{31}$$

$$\frac{\theta}{1+\theta}\gamma r f'_-(r) \ge A_0 f(r), \quad \forall r \in [0,R] \tag{32}$$

Proof We prove this by dividing $(\mathsf{G}\times\mathfrak{g})^2$ in to 3 regions (Fig. 1).

Region I: $\|\mu\| \le \gamma R$ and $\|Q\| \ge \epsilon$ (Reflection coupling) (Corresponding to Condition Eq. (29))

rc $\equiv 1$ in this case and we have

$$
\begin{aligned}
K_t &= \left[cf(r_t) + L\gamma^{-1}\left((1+\theta)\|Q_t\| - \theta\|Z_t\|\right)f'_-(r_t) + 4\gamma^{-1}f''(r_t)\right]\left(1 + \beta\|\mu_t\|^2\right) \\
&\quad + \beta f(r_t)\left[-2\gamma\|\mu_t\|^2 + 2L\|\mu_t\|\|Z_t\| + 8\gamma\|\mu_t\|^2\right] + 16\beta\|\mu_t\|f'_-(r_t) \\
&\le \left(c + \beta\left[2L\|\mu_t\|\|Z_t\| + 6\gamma\|\mu_t\|^2\right]f(r_t)\right)\left(1 + \beta\|\mu_t\|^2\right) \\
&\quad + \left(L\gamma^{-1}\left((1+\theta)\|Q_t\| - \theta\|Z_t\|\right) + 16\beta\|\mu_t\|\right)f'_-(r_t)\left(1 + \beta\|\mu_t\|^2\right) \\
&\quad + 4\gamma^{-1}f''(r_t)\left(1 + \beta\|\mu_t\|^2\right)
\end{aligned}
$$

$\|Z\| \le D$ and $\|\mu\| \le \gamma R$ give us

$$c + \beta \left[2L\|\mu_t\|\|Z_t\| + 6\gamma\|\mu_t\|^2 \right] \le c + \beta \left[2L\gamma RD + 6\gamma(\gamma R)^2 \right]$$

By the fact $\|\mu\| \le \gamma \max\{1, \alpha^{-1}\}r$ and $\|Q\| \le r$

$$
\begin{aligned}
&L\gamma^{-1}\left((1+\theta)\|Q\| - \theta\|Z\|\right) + 16\beta\|\mu\| \\
&\le L\gamma^{-1}\|Q\| + 16\beta\|\mu\| \\
&\le L\gamma^{-1}r + 16\beta\gamma\max\{1, \alpha^{-1}\}r
\end{aligned}
$$

As a result, by setting A_0, A_1, A_2 as in Eq. (30), we have

$$K_t \le \left[A_0 f(r) + A_1 r_t f'_-(r_t) + A_2 f''(r_t) \right]\left(1 + \beta\|\mu\|^2\right)$$

When Eq. (29) is satisfied, $K_t \le 0$ in this region.

Region II & III: $\|\mu\| \ge \gamma R + \epsilon$ (Synchronous coupling) (Corresponding to Condition Eq. (31))

In the case, we have $r \ge R$, which means $f'_-(r) = f''(r) = 0$.

$$
\begin{aligned}
K_t &= cf(r_t)\left(1 + \beta\|\mu_t\|^2\right) + \beta f(r_t)\left[-2\gamma\|\mu_t\|^2 + 2L\|\mu_t\|\|Z_t\|\right] \\
&\le f(r_t)\left[c - \beta(2\gamma - c - 2LD\gamma^{-1}R^{-1})\|\mu_t\|^2\right]
\end{aligned}
$$

which is non-negative given condition Eq. (31).

Region IV: $\|\mu\| \le \gamma R$ and $\|Q\| < \epsilon$ or $\gamma R < \|\mu\| < \gamma R + \epsilon$ (Mix between synchronous and reflection coupling) (Corresponding to Condition Eq. (32))

$0 \le \mathrm{rc} \le 1$ in this case and we have

$$
\begin{aligned}
K_t &\le \mathrm{rc}(Z_t, \mu_t)^2 \left[A_0 f(r_t) + A_1 r_t f'_-(r_t) + A_2 f''(r_t) \right]\left(1 + \beta\|\mu_t\|^2\right) \\
&\quad + \left(1 - \mathrm{rc}(Z_t, \mu_t)^2\right)\left[\left(c - 2\beta\gamma\|\mu_t\|^2 + 2\beta L\|\mu_t\|\|Z_t\|\right)f(r_t) + \left(\alpha\gamma\|Q_t\| - \theta L\gamma^{-1}\|Z_t\|\right)f'_-(r_t)\right]\left(1 + \beta\|\mu_t\|^2\right) \\
&\le \left[\left(c - 2\beta\gamma\|\mu_t\|^2 + 2\beta L\|\mu_t\|\|Z_t\|\right)f(r_t) + \left(\alpha\gamma\|Q_t\| - \theta L\gamma^{-1}\|Z_t\|\right)f'_-(r_t)\right]\left(1 + \beta\|\mu_t\|^2\right) \\
&\le \left[A_0 f(r_t) + \left(\alpha\gamma\epsilon - \theta L\gamma^{-1}\|Z_t\|\right)f'_-(r_t)\right]\left(1 + \beta\|\mu_t\|^2\right) \\
&\le \left[A_0 f(r_t) - \alpha^{-1}\theta L\gamma^{-1}r_t f'_-(r_t)\right]\left(1 + \beta\|\mu_t\|^2\right) + \alpha\gamma\epsilon f'_-(r_t)\left(1 + \beta\|\mu_t\|^2\right)
\end{aligned}
$$

Since $f'_- \le 1$ gives the desired result.

∎

F.3. Choose the parameters

In order to make sure the conditions in Lemma 37 satisfied, the parameters α (θ), β, R, the contraction rate c and the function f are needed to be carefully selected. The idea of choosing parameters here is inspired by Eberle (2016); Eberle et al. (2019).

F.3.1. CHOOSE f

At first glance, choosing f seems to be the most difficult part since the condition Eq. (29) is not very intuitive and we do not have a parametrization for function f. So, we start by giving f an explicit expression heuristically.

Denote φ as

$$\varphi(s) := \exp\left(-\frac{A_1}{2A_2}s^2\right)$$

i.e., φ is the solution of ODE $A_1 r\varphi(r) + A_2\varphi'(r) = 0$. An intuition for this equation on φ is by imaging $\varphi = f'$ in Eq. (29) with the term $A_0 f$ disregarded.

Since the ODE for φ omitted the positive $A_0 f$ term, $f(r)$ cannot be simply set as $\int_0^r \varphi(s)\mathrm{d}s$. Instead, a correction term ψ is introduced and f has the following form:

$$f(r) = \int_0^{\min\{r,R\}} \varphi(s)\psi(s)\mathrm{d}s \tag{33}$$

$\psi \leq 1$ is a function to be determined later. By denoting

$$\Phi(r) := \int_0^r \varphi(s)ds \tag{34}$$

we have $f \leq \Phi$ and when $r \leq R$,

$$
\begin{aligned}
A_0 f + A_1 r f'_- + A_2 f'' &= A_0 f + A_1 r(\varphi\psi) + A_2(\varphi\psi' + \varphi'\psi) \\
&= A_0 f + A_2\varphi\psi' \\
&\leq A_0\Phi + A_2\varphi\psi'
\end{aligned}
$$

This shows that $A_0\Phi(r) + A_2\varphi(r)\psi'(r) \leq 0$ is a sufficient condition for Eq. (29), which inspires us to choose ψ as

$$\psi(r) = 1 - \frac{A_0}{A_2}\int_0^r \Phi(s)\varphi(s)^{-1}ds, \quad \forall r \in [0, R] \tag{35}$$

Now we have an expression for f, we need to find a set of parameters $\alpha(\theta)$, β, R and c s.t. Eq. (31), (32) is satisfied.

F.3.2. CHOOSE $\alpha(\theta)$, β, R

First, we focus on condition Eq. (31). We set $\lambda c = \beta(6\gamma^3 R^2 + 2\gamma LRD)$ where λ is a constant to be determined, i.e., A_0 is assumed to be of the same order as c. Under such assumption, Eq. (31) can be simplified as

$$
\begin{aligned}
&c - \beta(2\gamma - c - 2LD\gamma^{-1}R^{-1})\gamma^2 R^2 \leq 0 \\
\Leftrightarrow\ &\beta(6\gamma^3 R^2 + 2\gamma LRD) \leq \lambda\beta(2\gamma - c - 2LD\gamma^{-1}R^{-1})\gamma^2 R^2 \\
\Leftrightarrow\ &6\gamma^2 R + 2LD \leq \lambda(2\gamma^2 R - c\gamma R - 2LD) \\
\Leftrightarrow\ &[2(\lambda - 3)\gamma - c\lambda]\gamma R \geq 2(1 + \lambda)LD
\end{aligned}
$$

The left-hand side must be positive, which means $\lambda > 3$. After simply choosing $\lambda = 4$ and assuming

$$c \leq \gamma/4 \tag{36}$$

we can choose R as

$$R = 10LD\gamma^{-2} \tag{37}$$

to make Eq. (31) satisfied. After R is chosen, solving $4c = \beta(6\gamma^3 R^2 + 2\gamma LRD)$ gives an explicit expression for β:

$$\beta = \frac{1}{155}c\gamma L^{-2}D^{-2} \tag{38}$$

Before we proceed, we give φ a lower bound by adding more constraints to the parameters. The benefit is that we will have a better estimate for the conditions depending on φ (Φ) and f, i.e., Eq. (32). $\frac{A_1}{2A_2} = \frac{\alpha}{8}\gamma^2 + 2\beta\gamma^2 \max\{1, \alpha^{-1}\}$ and we have

$$\varphi(s) = \exp\left(-\frac{Ls^2}{8}(1+\theta) - 2\gamma^2\beta\max\{1, \alpha^{-1}\}s^2\right)$$

And we want to make sure

$$\varphi(s) \geq \exp\left(-2 - \frac{Ls^2}{8}\right), \quad \forall s \in [0, R] \tag{39}$$

Since this only need to be satisfied on $[0, R]$, it can be achieved by ensuring $2\beta\gamma^2 R^2 \max\{1, \alpha^{-1}\} \leq 1$ and $\theta LR^2/8 \leq 1$, which further leads to the following choice of θ

$$\theta = 8L^{-1}R^{-2} = 0.08L^{-3}D^{-2}\gamma^4 \tag{40}$$

and one more condition on c:

$$c \leq \frac{31}{40}\min\{\gamma, (1+\theta)L\gamma^{-1}\} \tag{41}$$

To summarize, comparing with the conditions in Lemma 37 (Eq. (29), (31), (32)), we made some artificial conditions and have a set of more explicit choice of constants (Eq. (37), (38), (40)) with conditions (Eq. (32), (36), (41)).

F.3.3. CHOOSE c

Now, the last part is to choose a suitable parameter c. We use all the choice of parameters mentioned earlier in Sec. F.3.1 and F.3.2. Given Lemma 37, we only need to check Eq. (29), (31), (32) are satisfied.

Lemma 38 *By our choice of f (Eq. (33)), θ (Eq. (40)), β (Eq. (38)), R (Eq. (37)), we have conditions in Lemma 37 ((29), (31), (32)) are satisfied for*

$$c \leq c_* := \min\left\{\frac{\gamma}{4}, \frac{31}{40}(1+\theta)L\gamma^{-1}, \frac{L^{3/2}R}{5\sqrt{2\pi}\gamma}e^{-\frac{LR^2}{8}}, \frac{\theta\gamma}{5(1+\theta)}\min\left\{e^{-\frac{5}{2}}, \frac{\sqrt{L}R}{\sqrt{2\pi}}e^{-\frac{LR^2}{8}}\right\}\right\} \tag{42}$$

Proof [Proof of Lemma 38] Condition Eq. (29) is satisfied by our choice of f in Eq. (33).

Eq. (31) is satisfied by Eq. (36), with extra condition Eq. (36).

The rest of the proof will be devoted to finding an appropriate c for Eq. (32).

We first ensure a lower bound by $\psi \geq 1/2$. By the explicit expression of ψ in Eq. (35), this is equal to

$$\frac{2A_2}{A_0} \leq \int_0^R \Phi(s)\varphi(s)^{-1}\mathrm{d}s \tag{43}$$

Our choice of β gives $A_0 = 5c$. Consequently, Eq. (43) is satisfied when

$$c \leq \frac{8}{5}\gamma^{-1} / \int_0^R \Phi(s)\varphi(s)^{-1}\mathrm{d}s$$

Since our the property Eq. (39) is guaranteed by our choice of β and extra condition on c in Eq. (41), we have the following upper bound of Φ from the definition of Φ in Eq. (34)

$$\Phi \leq \int_0^\infty \exp(-2 - Ls^2/8)\mathrm{d}s = e^{-2}\sqrt{2\pi/L}$$

This upper bound of Φ gives an upper bound of $\int_0^R \Phi(s)\varphi(s)^{-1}\mathrm{d}s$:

$$\int_0^R \Phi(s)\varphi(s)^{-1}\mathrm{d}s$$
$$\leq \Phi(R)\int_0^R \varphi(s)^{-1}\mathrm{d}s$$
$$\leq \Phi(R)e^2 \int_0^R \exp(Ls^2/8)\mathrm{d}s$$
$$\leq \Phi(R)e^2 \frac{8}{LR}\exp(LR^2/8)$$
$$\leq \sqrt{2\pi/L}\frac{8}{LR}\exp(LR^2/8)$$

In the end, an sufficient condition for Eq. (43) is

$$c \leq \frac{L^{3/2}R}{5\sqrt{2\pi}\gamma}\exp(-LR^2/8) \tag{44}$$

By our choice of θ in Eq. (40) and apply condition Eq. (41), we have Eq. (39).

Next, we are ready to make Eq. (32) satisfied. By the expression for f, we have $f \leq \Phi$. Using the property $\psi \geq 1/2$ guaranteed by Eq. (44), we have $f' = \varphi\psi \geq \frac{1}{2}\varphi$. Together with $A_0 = 5c$ and Eq. (39), we know Eq. (32) is satisfied given

$$\frac{5c(1+\theta)}{\theta\gamma} \leq \inf_{r\in[0,R]} \frac{r\varphi(r)}{\Phi(r)}$$

Now we estimate a lower bound for $\inf_{r\in[0,R]} \frac{r\varphi(r)}{\Phi(r)}$. First we notice that

$$\frac{r\varphi(r)}{\Phi(r)} \geq \frac{r\exp(-2 - Lr^2/8)}{\int_0^r \exp(-2 - Ls^2/8)ds} = \frac{r\exp(-Lr^2/8)}{\int_0^r \exp(-Ls^2/8)ds}$$

For $r \geq \frac{2}{\sqrt{L}}$, we have $\frac{\mathrm{d}}{\mathrm{d}r} r \exp(-Lr^2/8) = (1-Lr^2/4)\exp(-Lr^2/8) \leq 0$ and $\frac{r\varphi(r)}{\Phi(r)} \geq \frac{R\exp(-LR^2/8)}{\int_0^R \exp(-Ls^2/8)ds} \geq R\exp(-LR^2/8)/\sqrt{2\pi/L}$.

For $r \leq \frac{2}{\sqrt{L}}$, we have $r/\Phi \leq 1$ and $\frac{r\varphi(r)}{\Phi(r)} \geq \varphi(r) \geq \exp(-5/2)$.

As a result, a sufficient condition for Eq. (32) is

$$c \leq \frac{\theta\gamma}{5(1+\theta)} \min\left\{\exp(-5/2), R\exp(-LR^2/8)/\sqrt{2\pi/L}\right\} \tag{45}$$

Since we added extra conditions on c in Eq. (36), (41)), a summary of all the conditions on c, Eq. (36), (41), (44), (45) give us the desired result. ∎

F.4. More discussion on the convergence rate c

Here we discuss the order of c_*, the upper bound of convergence rate defined in Eq. (42). By denoting

$$\Lambda := \frac{LR^2}{8} \sim L^3 D^2 \gamma^{-4} \tag{46}$$

We have an estimate of the order of c_*:

$$c_* \sim \min\left\{\gamma, L\gamma^{-1}, L\gamma^{-1}\sqrt{\Lambda}e^{-\Lambda}, \gamma\Lambda^{-1}, \gamma\Lambda^{-\frac{1}{2}}e^{-\Lambda}\right\} \tag{47}$$

Notice that the inner product in Lemma 3 can be rescaled $\langle\cdot,\cdot\rangle \to a\langle\cdot,\cdot\rangle$, and in this case $L \to a^{-2}L$ and $D \to aD$, but the value LD^2 will remain the same. As a result, we consider dependence of the convergence rate c_* on γ and LD^2 when $LD^2 \to \infty$:

- For the critical choice $\gamma \sim \sqrt{L}$, we have $c_* \sim D^{-\frac{1}{2}}e^{-LD^2}$. This is also the common choice in the Euclidean case.

- For the underdamped choice $\gamma \lesssim \sqrt{L}$, we have $c_* \sim \gamma\min\left\{\Lambda^{-1}, \Lambda^{-\frac{1}{2}}e^{-\Lambda}\right\}$, which is slower than the kinetic choice when $LD^2 \to \infty$.

- For the overdamped choice $\gamma \gtrsim \sqrt{L}$, it is hard to simplify the expression for c_*. One may suggest $\gamma \sim (L^3D^2)^{\frac{1}{4}}$ to make $\Lambda \sim 1$. In this case, we have $c_* \sim L\gamma^{-1} \sim D^{\frac{1}{2}}L^{-\frac{1}{4}}$. However, the research in Euclidean kinetic Langevin suggest this will lead to the requirement of small step size and will not remove the exponential decay of convergence rate when D is large.

In summary, our recommendation is $\gamma \sim \sqrt{L}$, and we have the convergence rate c decays exponentially when D (diameter of Lie group) grows. One may worry that our rate is slow, but it actually should be slow because we only assumed compactness and smooth log density, and metastability can exist: without convexity conditions on U, the potential barriers between the potential wells of the non-convex potential function U stops the Langevin dynamics from mixing fast. We can see that by comparing, for example, our rate with some existing results by Gaussian mixture models (GMM).

We consider the mixture of 2 Gaussians $N(x, \sigma I)$ and $N(y, \sigma I)$ with equal weight in Euclidean space. Schlichting (2019) provides an estimate of LSI constant ϱ for this GMM model showing an exponential $\varrho \gtrsim \exp(-\|x - y\|^2/\sigma)$, and Ma et al. (2021) showing a convergence rate ϱ. This means the result Ma et al. (2021) has a convergence rate exponentially decaying w.r.t. the distance between 2 Gaussian models. In other words, exponential long time is required to jump between potential well when they are far, and since we have no convexity assumption on potential function, this slow rate cannot be avoided.

Our choice of γ and the quantification of convergence rate are consistent with the Euclidean case (Eberle et al., 2019, Sec. 2.6).

Appendix G. Proof for Thm. 7

Proof [Proof for Thm. 7] By Cor. 35, we have both the density of (g_t, ξ_t) and $(\hat{g}_t, \hat{\xi}_t)$ are absolute continuous w.r.t. $\mathrm{d}g\mathrm{d}\xi$ for any $t \in [0, \infty)$, which gives us $\mathbb{E}_{(g_t, \xi_t, \hat{g}_t, \hat{\xi}_t)} \mathbb{1}_{\{\hat{g}_t^{-1} g_t \in N\}} \delta_t \equiv 0, \forall t \in [0, \infty)$.

As a result, by taking expectation to Eq. (11) and choosing the parameters α, β and R as stated in Lemma 38, we have

$$\mathbb{E}\rho_t \leq e^{-ct}\mathbb{E}\rho_0 + \epsilon\alpha\gamma \int_0^t \mathbb{E}\left(1 + \beta\|\mu_s\|^2\right) \mathrm{d}s$$

By the fact that $\mathbb{E}\|\mu_s\|^2$ is upper bounded by $4C_2$ (Lemma 46) since $\|\mu_s\|^2 \leq 2\|\xi_s\|^2 + 2\|\hat{\xi}_s\|^2$, we can take the limit when $\epsilon \to 0$ and have

$$\mathbb{E}\rho_t \leq e^{-ct}\mathbb{E}\rho_0$$

By choosing the optimal joint distribution for initial distribution $\nu_0 = \mathrm{Law}(g_0, \xi_0)$ and $\hat{\nu}_0 = \mathrm{Law}(\hat{g}_0, \hat{\xi}_0)$, we have

$$\begin{aligned}
&W_\rho(\nu_0, \hat{\nu}_0) \\
&= \mathbb{E}\rho((g_0, \xi_0), (\hat{g}_0, \hat{\xi}_0)) \\
&\geq e^{ct}\mathbb{E}\rho((g_t, \xi_t), (\hat{g}_t, \hat{\xi}_t)) \\
&\geq e^{ct}W_\rho(\mathrm{Law}(g_t, \xi_t), \mathrm{Law}(\hat{g}_t, \hat{\xi}_t)) \\
&= e^{ct}W_\rho(\nu_t, \hat{\nu}_t)
\end{aligned}$$

$\blacksquare$

Proof [Proof of Lemma 4] We consider 2 cases separately: 1) $\|\xi - \hat{\xi}\| \geq 2D$; 2) $\|\xi - \hat{\xi}\| < 2D$, where D is the diameter of the Lie group.

Case 1: $\|\xi - \hat{\xi}\| \geq 2D$. In this case, we have

$$\|\xi - \hat{\xi}\|^2 \geq \frac{1}{2}\|\xi - \hat{\xi}\|^2 + \frac{1}{2}D^2 \geq \frac{1}{2}\|\xi - \hat{\xi}\|^2 + \frac{1}{2}d^2(g, \hat{g})$$

The definition of r in Eq. (8) gives

$$
\begin{aligned}
&r((g,\xi),(\hat{g},\hat{\xi})) \\
&\geq \min\{\alpha,1\}\left(d(g,\hat{g}) + \left\|\log \hat{g}^{-1}g + \gamma^{-1}(\hat{\xi}-\xi)\right\|\right) \\
&\geq \min\{\alpha,1\}\gamma^{-1}\left\|\hat{\xi}-\xi\right\| \\
&\geq 2\min\{\alpha,1\}\gamma^{-1}D
\end{aligned}
$$

which infers by the definition of ρ,

$$
\begin{aligned}
&\rho((g,\xi),(\hat{g},\hat{\xi})) \\
&\geq f(2\min\{\alpha,1\}\gamma^{-1}D)(1 + \beta d^2((g,\xi),(\hat{g},\hat{\xi}))/2) \\
&\geq \frac{\beta}{2}f(2\min\{\alpha,1\}\gamma^{-1}D)d^2((g,\xi),(\hat{g},\hat{\xi}))
\end{aligned}
$$

Case 2: $\left\|\xi-\hat{\xi}\right\| < 2D$. In this case, $r \leq (\alpha+1+2\gamma^{-1})D$. By the concavity of f, we have

$$
f(r) \geq \frac{f((\alpha+1+2\gamma^{-1})D)}{(\alpha+1+2\gamma^{-1})D}r \geq \frac{f((\alpha+1+2\gamma^{-1})D)}{((\alpha+1+\gamma^{-1})^2D^2}r^2
$$

Since

$$
\begin{aligned}
&r((g,\xi),(\hat{g},\hat{\xi})) \\
&= \frac{\alpha}{2}d(g,\hat{g}) + \frac{\alpha}{2}d(g,\hat{g}) + \left\|\log \hat{g}^{-1}g + \gamma^{-1}(\hat{\xi}-\xi)\right\| \\
&\geq \frac{\alpha}{2}d(g,\hat{g}) + \gamma^{-1}\min\{1,\alpha/2\}\left\|\xi-\hat{\xi}\right\|
\end{aligned}
$$

we have

$$
\rho \geq f(r) \geq \frac{f((\alpha+1+2\gamma^{-1})D)}{(\alpha+1+\gamma^{-1})^2D^2}r^2 \geq \frac{f((\alpha+1+2\gamma^{-1})D)}{(\alpha+1+\gamma^{-1})^2D^2}\min\left\{\frac{\alpha^2}{4},\frac{1}{\gamma^2},\frac{\alpha^2}{4\gamma^2}\right\}d^2
$$

By defining C_ρ as

$$
C_\rho = \min\left\{\frac{\beta}{2}f(2\min\{\alpha,1\}\gamma^{-1}D), \frac{f((\alpha+1+2\gamma^{-1})D)}{(\alpha+1+\gamma^{-1})^2D^2}\min\left\{\frac{\alpha^2}{4},\frac{1}{\gamma^2},\frac{\alpha^2}{4\gamma^2}\right\}\right\} \tag{48}
$$

we have $C_\rho d^2 \leq \rho$. ∎

From Lemma 4, we directly have the following corollary, which bounds W_2 by W_ρ, and further gives us convergence under W_2 (Thm. 8).

Corollary 39 (Control of W_2 by W_ρ) *For any distributions ν and $\hat{\nu}$ on $\mathsf{G} \times \mathfrak{g}$, we have*

$$
C_\rho W_2^2(\nu,\hat{\nu}) \leq W_\rho(\nu,\hat{\nu})
$$

Proof [Proof of Cor. 39] Suppose $\pi \in \Pi(\nu, \hat{\nu})$ is the optimal coupling between ν and $\hat{\nu}$ under ρ, i.e., $\int \rho(x,y)d\gamma = W_\rho(\nu, \hat{\nu})$. Lemma 4 gives

$$
\begin{aligned}
W_\rho(\nu, \hat{\nu}) &= \int \rho \mathrm{d}\pi \\
&\geq \int C_\rho d^2 \mathrm{d}\pi \\
&\geq C_\rho W_2^2(\nu, \hat{\nu})
\end{aligned}
$$

∎

Proof [Proof of Thm .8] This is a direct corollary for Thm. 7 and Cor. 39. ∎

Appendix H. More details about Sec. 5

H.1. More discussion on the splitting discretization

Splitting discretization is less commonly used when discretizing kinetic Langevin SDE in the Euclidean space. This is because another discretization, which we will refer to as the 'exponential integrator discretization' (Eq.49) is easier to analyze. However, that exponential integrator discretization works in Euclidean space but not on Lie groups.

More precisely, in the existing literature studying kinetic Langevin in Euclidean spaces, time discretization of SDE Eq. (17) is commonly done by the explicit solution of

$$
\begin{cases}
\dot{q} = p \\
dp = -\gamma p dt - \nabla U(q_0)dt + \sqrt{2\gamma}dW
\end{cases}
\tag{49}
$$

i.e., by viewing ∇U as constant and using a closed-form solution for the rest. The only numerical error it introduces is the error in the gradient of the potential function. This special discretization also is of key importance in many proofs, e.g., Ma et al. (2021); Altschuler and Chewi (2023).

However, although in the Euclidean case, it is a linear SDE and admits a closed-form solution, in the Lie group case it becomes

$$
\begin{cases}
\dot{g} = T_e \mathsf{L}_g \xi \\
\mathrm{d}\xi = -\gamma \xi dt - T_g \mathsf{L}_{g^{-1}} \nabla U(g_0)dt + \sqrt{2\gamma}dW
\end{cases}
$$

It is no longer a linear SDE due to the involvement of the operator $T_e \mathsf{L}_g$, and this SDE does not have a closed-form solution, hence not implementable as a numerical algorithm.

To see this more clearly, let's consider the matrix group case in Example 1. In this case, $T_e \mathsf{L}_g \xi = g\xi$, where both g and ξ are $n \times n$ matrices and $g\xi$ is the matrix multiplication. As a result, even though the ξ equation can be analytically solved, resulting in a precise expression ξ as a function of time, the equation $\dot{g} = g\xi$ is a non-autonomous linear system with coefficient (ξ) not necessarily commuting at different times, and hence g does not admit a closed-form solution.

Due to the nonlinearity of the $g\xi$ term, we use the splitting discretization Eq. (14) instead, where we step g and ξ separately. In that case, when we evolve $\dot{g} = g\xi$ in the matrix

45

group, we view ξ as constant and that is a linear equation in g, which admits a closed-form solution. However, extra numerical errors other than the error in the gradient of potential may be introduced compared to the exponential integrator discretization Eq. (49), which makes our analysis more non-trivial.

At the end of this section, we give proof for the property of structure-preserving for splitting discretization Thm. 9.

Proof [of Thm. 9]. Our numerical scheme is given by alternatively evolving the two SDEs. Suppose (g_k, ξ_k) preserves the manifold structure, i.e., $g \in \mathsf{G}$ and $\xi \in \mathfrak{g}$. For the first step when updating ξ by evolving Eq.12, since it is an SDE in $\mathfrak{g}$, which is a linear space, we have $\xi_{k+1} \in \mathfrak{g}$. For the second step when updating g by evolving Eq. (13), since the group exponential is always in the group, i.e., $\exp(h\xi_{k+1}) \in \mathsf{G}$, we have $g_{k+1} = g_k \exp(h\xi_{k+1})$ is in G since the group is closed under multiplication.

By induction, We have if $g_0 \in \mathsf{G}$ and $\xi_0 \in \mathfrak{g}$, the group structure is preserved for the whole trajectory, i.e., $g_k \in \mathsf{G}$ and $\xi_k \in \mathfrak{g}$ for any k. ∎

Remark 40 (Example of structure preserving on $\mathsf{SO}(n)$) *When we use matrix representation for $\mathsf{SO}(n)$ (Example 2), we have the Lie group structure is $X^\top X = I$, $\xi + \xi^\top = 0$. As a result, Thm. 9 tell us $X_k^\top X_k = I$, $\xi_k + \xi_k^\top = 0$ for all k under any step size h.*

Remark 41 (Comparison with (Cheng et al., 2022)) *In a brilliant paper by Cheng et al. (2022), a quantification of the discretization of the Langevin on the manifold (without momentum) is proposed. However, our Lie group has the group structure and our error bound can be more explicit. What's more, our discretization is not simply geometric Euler-Maruyama discretization but a two-stage splitting. How to expand the bound for numerical error for exponential integrator in their work to splitting discretization is unclear yet.*

H.2. Proof of Thm. 10

The proof of Thm. 10 will be in several steps: 1) we find a shadow SDE whose distribution at t is the distribution of our numerical method evolve one step with step size t (Lemma 42); 2) we upper bound some terms in Lemma 43, 44, 45, 46; 3) using those estimations, we derive an ODE whose solution at time t is the upper bound of one step numerical error of our numerical scheme (Thm. 10).

Lemma 42 (Exact SDE for splitting scheme) *Denote $(\tilde{g}_h, \tilde{\xi}_h)$ as the one-step evolution of our splitting numerical scheme Eq. (14) with step size t starting from (g_0, ξ_0). Then $(\tilde{g}_h, \tilde{\xi}_h)$ is the solution of the following SDE at time h:*

$$\begin{cases} \mathrm{d}\tilde{g}_t = T_e\mathsf{L}_{\tilde{g}_t}\left(\tilde{\xi}_t\mathrm{d}t + t\frac{1-e^{\mathrm{ad}_{t\tilde{\xi}_t}}}{\mathrm{ad}_{t\tilde{\xi}_t}}\left(-\gamma\tilde{\xi}_t\mathrm{d}t - T_{g_0}\mathsf{L}_{g_0^{-1}}\nabla U(\tilde{g}_0)\mathrm{d}t + \sqrt{2\gamma}\mathrm{d}W_t\right) + \gamma t^2 \sum_{ij} \mathrm{d}^2\exp(\tilde{\xi}_t)(e_i, e_j)\mathrm{d}t\right) \\ \mathrm{d}\tilde{\xi}_t = -\gamma\tilde{\xi}_t\mathrm{d}t - T_{g_0}\mathsf{L}_{g_0^{-1}}\nabla U(\tilde{g}_0)\mathrm{d}t + \sqrt{2\gamma}\mathrm{d}W_t \end{cases}$$
(50)

where $\mathrm{d}^2\exp : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ is defined as

$$\mathrm{d}^2\exp(x)(\xi, \hat{\xi}) = \lim_{h\to 0}\left\langle \mathrm{d}\exp(x+h\xi), \hat{\xi}\right\rangle - \left\langle \mathrm{d}\exp(x), \hat{\xi}\right\rangle$$
(51)

Proof The SDE for ξ is from the definition of the numerical discretization directly. We focus on the SDE for g. The update of g gives $\tilde{g}_t = g \exp(t\tilde{\xi}_t)$, and we will first write down the SDE $t\tilde{\xi}_t$ satisfied as the following:

$$\mathrm{d}t\tilde{\xi}_t = (1 - t\gamma)\tilde{\xi}_t\mathrm{d}t - tT_{g_0}\mathsf{L}_{g_0^{-1}}\nabla U(\tilde{g}_0)\mathrm{d}t + t\sqrt{2\gamma}\mathrm{d}W_t$$

By Ito's formula and the expression of $\mathrm{d}\exp$ in Eq. (24),

$$\mathrm{d}\tilde{g}_t = \mathrm{d}\exp(t\tilde{\xi}_t)$$

$$= T_e\mathsf{L}_{\exp(t\tilde{\xi}_t)}\left(\mathrm{d}\exp_{t\tilde{\xi}_t}(\mathrm{d}t\tilde{\xi}_t) + \gamma t^2\sum_{ij}\mathrm{d}^2\exp(t\tilde{\xi}_t)(e_i, e_j)\mathrm{d}t\right)$$

where the second term in the bracket is Ito's correction drift term. ∎

Lemma 43 (Upper bound for $\mathrm{d}^2\exp$) *The map $\mathrm{d}^2\exp(X) : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ defined in Eq. (51) has upper bound C, i.e.,*

$$\left\|\mathrm{d}^2\exp(X)\right\|_{op} \leq C$$

where C is defined in Eq. (21) and $\|\cdot\|_{op}$ is the operator norm (maximum eigenvalue of a linear operator).

Proof [Proof of Lemma 43] By the definition of $\mathrm{d}^2\exp$ given by Eq. (51) and $\mathrm{d}\exp$ in Eq. 24, we have

$$\mathrm{d}^2\exp(X)(Y, Z) = \lim_{h\to 0}\langle\mathrm{d}\exp(X + hY), Z\rangle - \langle\mathrm{d}\exp(x), Z\rangle$$

$$= \lim_{h\to 0}\frac{1 - \exp(-\mathrm{ad}_{X+hY})}{\mathrm{ad}_{X+hY}}Z - \frac{1 - \exp(-\mathrm{ad}_X)}{\mathrm{ad}_x}Z$$

Setting $Y = Z$,

$$\left\|\mathrm{d}^2\exp(X)(Y, Y)\right\| \leq \|\mathrm{ad}_Y\|_{op}\|Y\|\max\left\{\frac{\mathrm{d}}{\mathrm{d}x}\mathfrak{R}\frac{1 - \exp(xi)}{xi}\right\} \leq C\|Y\|^2$$

∎

Lemma 44 (Upper bound $\frac{\partial^2}{\partial e_i\partial e_j}d^2(\cdot, g)$) *Define*

$$\frac{\partial^2}{\partial e_i\partial e_j}d^2(\tilde{g}, g) := \lim_{h\to 0}\frac{d^2(\tilde{g}\exp(he_i)\exp(he_j), g\exp(he_i)) - d^2(\tilde{g}\exp(he_j), g)}{h^2}$$

Then we have

$$\frac{\partial^2}{\partial e_i\partial e_j}d^2(\tilde{g}, g) \leq Cd(\tilde{g}, g)$$

47

Proof [Proof of Lemma 44]

$$\frac{\partial^2}{\partial e_i \partial e_j} d^2(\tilde{g}, g) := \lim_{h \to 0} \frac{d^2(\tilde{g}\exp(he_i)\exp(he_j), g\exp(he_i)) - d^2(\tilde{g}\exp(he_j), g)}{h^2}$$
$$= \lim_{h \to 0} \frac{d^2(\tilde{g}\exp(he_i)\exp(he_j)\exp(-he_j), g) - d^2(\tilde{g}, g)}{h^2}$$
$$= \lim_{h \to 0} \frac{d^2(\tilde{g}\exp(h^2[e_i, e_j]), g) - d^2(\tilde{g}, g)}{h^2}$$
$$\leq 2d(\tilde{g}, g)\|[e_i, e_j]\|$$
$$\leq 2Cd(\tilde{g}, g)$$

■

Lemma 45 *The operator norm for* $\mathrm{d}\exp$ *is bounded by*

$$\|\mathrm{d}\exp_X\|_{op} \leq 1$$

Proof [Proof of Lemma 45] This can be derived from Eq.(24). Since $\left\|\frac{1-e^{xi}}{xi}\right\| \leq 1, \forall x$ and all the eigenvalue of ad_X is pure imaginary by the skew-symmetricity of ad_X operator, we have the desired result. ■

Lemma 46 *Suppose we initialize* $\xi(0) \equiv 0$ *and we have* $\|\nabla f\|$ *is upper bounded by* LD. *Then we have* $\mathbb{E}\|\xi\|^k < \infty$ *for both SDE Eq. (4) and splitting scheme Eq. (14).*

Proof [Proof of Lemma 46] The drift term for momentum has magnitude $\leq LD$ and by Tanaka formula, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\|\xi\| = -\gamma\mathbb{E}\|\xi\| - \mathbb{E}\left\langle\frac{\xi}{\|\xi\|}, \nabla f\right\rangle$$
$$\leq -\gamma\mathbb{E}\|\xi\| + LD$$

as a result, $\mathbb{E}\|\xi\| \leq \frac{LD}{\gamma} := C_1$.

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\|\xi\|^2 = -2\gamma\mathbb{E}\|\xi\|^2 - 2\mathbb{E}\langle\xi, \nabla f\rangle + m$$
$$\leq -2\gamma\mathbb{E}\|\xi\|^2 + 2LDC_1 + m$$

As a result,

$$\mathbb{E}\|\xi\|^2 \leq \frac{2LDC_1 + m}{2\gamma} := C_2$$

We prove by induction. Suppose for $i = 1, \ldots k-1$ $\mathbb{E}\|\xi\|^i \leq C_i$, we have for $k \geq 3$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\|\xi\|^k = -\gamma k\mathbb{E}\|\xi\|^k - k\mathbb{E}\|\xi\|^{k-2}\langle\xi, \nabla f\rangle + \gamma\mathbb{E}\Delta\|\xi\|^k$$
$$= -\gamma k\mathbb{E}\|\xi\|^k - k\mathbb{E}\|\xi\|^{k-2}\langle\xi, \nabla f\rangle + \gamma k(k-1)\mathbb{E}\|\xi\|^{k-2}$$
$$\leq -\gamma k\mathbb{E}\|\xi\|^k + kLDC_{k-1} + \gamma k(k-1)C_{k-2}$$

Since we have $\xi(0) \equiv 0$, we have $\mathbb{E}\|\xi\|^k \leq \frac{LDC_{k-1}}{\gamma} + (k-1)C_{k-2} := C_k < \infty$ ∎

Proof [Proof of Thm. 10] We consider the evolution of $\mathbb{E}d^2((g_t, \xi_t), (\tilde{g}_t, \tilde{\xi}_t))$, where (g_t, ξ_t) and $(\tilde{g}_t, \tilde{\xi}_t)$ are 2 trajectories strating from the same point (g_0, ξ_0).

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}d^2((g_t, \xi_t), (\tilde{g}_t, \tilde{\xi}_t))$$
$$= \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}d^2(g_t, \tilde{g}_t) + \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}d^2(\xi_t, \tilde{\xi}_t)$$

We quantify $\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}d^2(\xi_t, \tilde{\xi}_t)$ first:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\|\tilde{\xi}_t - \xi_t\|^2 \leq 2\mathbb{E}\left[\|\tilde{\xi}_t - \xi_t\|\|\nabla U(g_t) - \nabla(g_0)\|\right]$$
$$\leq 2L\mathbb{E}\left[\|\tilde{\xi}_t - \xi_t\|d(g_t, g_0)\right]$$
$$\leq 2L\mathbb{E}\left[\|\tilde{\xi}_t - \xi_t\|\int_0^t \|\xi_t\|\mathrm{d}t\right]$$
$$\leq L\mathbb{E}\|\tilde{\xi}_t - \xi_t\|^2 + L\mathbb{E}\left(\int_0^t \|\xi_t\|\mathrm{d}t\right)^2$$
$$\leq L\mathbb{E}\|\tilde{\xi}_t - \xi_t\|^2 + t^2 LC_2$$

where the last line is because Cauchy-Schwarz inequality and Lemma 46 gives us

$$\mathbb{E}\left(\int_0^t \|\xi_t\|\mathrm{d}t\right)^2 \leq t\int_0^t \mathbb{E}\|\xi_t\|^2\mathrm{d}t \leq t^2 C_2$$

Gronwall's inequality gives $\mathbb{E}\|\tilde{\xi}_t - \xi_t\|^2 \leq x$ where x is the solution of ODE $\dot{x} = Lx + C_2 Lt^2$ with initial condition $x(0) = 0$.

Next, we bound $\mathbb{E}d^2(\tilde{g}_t, g_t)$. When we have $\tilde{g}_t^{-1}g_t \notin N$, Ito's formula gives

$$\mathrm{d}d^2(\tilde{g}_t, g_t)$$
$$= 2\left\langle \log g_t^{-1}\tilde{g}_t, \mathrm{d}\tilde{g}_t - \mathrm{d}g_t\right\rangle\mathrm{d}t + \gamma t^2 \sum_{ijk}\left\langle\frac{1 - e^{\mathrm{ad}_{t\tilde{\xi}_t}}}{\mathrm{ad}_{t\tilde{\xi}_t}}e_k, e_i\right\rangle\left\langle\frac{1 - e^{\mathrm{ad}_{t\tilde{\xi}_t}}}{\mathrm{ad}_{t\tilde{\xi}_t}}e_k, e_j\right\rangle\frac{\partial^2}{\partial e_i \partial e_j}d^2(\tilde{g}_t, g_t)\mathrm{d}t$$

where the second term is the Ito's correction. It can be bounded by Lemma 44 and 45.

$$\sum_{ijk}\left\langle\frac{1 - e^{\mathrm{ad}_{t\tilde{\xi}_t}}}{\mathrm{ad}_{t\tilde{\xi}_t}}e_k, e_i\right\rangle\left\langle\frac{1 - e^{\mathrm{ad}_{t\tilde{\xi}_t}}}{\mathrm{ad}_{t\tilde{\xi}_t}}e_k, e_j\right\rangle\frac{\partial^2}{\partial e_i \partial e_j}d^2(\tilde{g}_t, g_t) \leq 2Cm^3 d(\tilde{g}_t, g_t)$$

After taking expectations, we only need to consider $\tilde{g}_t^{-1}g_t \in \mathsf{G}\backslash N$ since N is zero measured (Cor. 35). Also, the local martingale part can be eliminated. Using Lemma 43 and 46 we

have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}d^2(\tilde{g}_t,g_t) = \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{\tilde{g}_t^{-1}g_t\notin N}d^2(\tilde{g}_t,g_t)$$

$$= 2\mathbb{E}\left\langle \log g_t^{-1}\tilde{g}_t, \tilde{\xi}_t - \xi_t + t\frac{1-e^{\mathrm{ad}_{t\tilde{\xi}_t}}}{\mathrm{ad}_{t\tilde{\xi}_t}}\left(-\gamma\tilde{\xi}_t - T_{g_0}\mathsf{L}_{g_0^{-1}}\nabla U(\tilde{g}_0)\right) + \gamma t^2\sum_{ij}\mathrm{d}^2\exp(t\tilde{\xi}_t)(e_i,e_j)\right\rangle$$

$$+ \gamma t^2\sum_{ijk}\left\langle\frac{1-e^{\mathrm{ad}_{t\tilde{\xi}_t}}}{\mathrm{ad}_{t\tilde{\xi}_t}}e_k,e_i\right\rangle\left\langle\frac{1-e^{\mathrm{ad}_{t\tilde{\xi}_t}}}{\mathrm{ad}_{t\tilde{\xi}_t}}e_k,e_j\right\rangle\frac{\partial^2}{\partial e_i\partial e_j}d^2(\tilde{g}_t,g_t)$$

$$\leq \mathbb{E}d^2(\tilde{g}_t,g_t) + \mathbb{E}\left\|\tilde{\xi}_t - \xi_t + t\frac{1-e^{\mathrm{ad}_{t\tilde{\xi}_t}}}{\mathrm{ad}_{t\tilde{\xi}_t}}\left(-\gamma\tilde{\xi}_t - T_{g_0}\mathsf{L}_{g_0^{-1}}\nabla U(\tilde{g}_0)\right) + \gamma t^2\sum_{ij}\mathrm{d}^2\exp(t\tilde{\xi}_t)(e_i,e_j)\right\|^2$$

$$+ 2\gamma t^2 Cm^3\mathbb{E}d(\tilde{g}_t,g_t)$$

$$\leq \mathbb{E}d^2(\tilde{g}_t,g_t) + 4\left(\mathbb{E}\|\tilde{\xi}_t - \xi_t\|^2 + \gamma^2 t^2\mathbb{E}\|\tilde{\xi}_t\|^2 + t^2\|\nabla U(\tilde{g}_0)\|^2 + \gamma^2 t^4 C^2 m^2\right)$$

$$+ \gamma t^2 Cm^3\mathbb{E}d^2(\tilde{g}_t,g_t) + \gamma t^2 Cm^3$$

$$\leq \mathbb{E}d^2(\tilde{g}_t,g_t) + 4\left(\mathbb{E}\|\tilde{\xi}_t - \xi_t\|^2 + \gamma^2 t^2 C_2 + t^2 LD + \gamma^2 t^4 C^2 m^2\right) + \gamma t^2 Cm^3\mathbb{E}d^2(\tilde{g}_t,g_t) + \gamma t^2 Cm^3$$

$$= \left(1 + \gamma Cm^3 t^2\right)\mathbb{E}d^2(\tilde{g}_t,g_t) + 4\mathbb{E}\|\tilde{\xi}_t - \xi_t\|^2 + \left(4\gamma^2 C_2 + 4LD + \gamma Cm^3\right)t^2 + 4\gamma^2 C^2 m^2 t^4$$

Gronwall's inequality gives $\mathbb{E}d^2(\tilde{g}_t,g_t) \leq y(t)$ where y is the solution of ODE

$$\dot{y} = \left(1 + \gamma Cm^3 t^2\right)y + 4x + \left(4\gamma^2 C_2 + 4LD + \gamma Cm^3\right)t^2 + 4\gamma^2 C^2 m^2 t^4$$

with initial condition $y(0) = 0$.

To summarize, the ODE for $x(t)$ and $t(t)$ are given by

$$\begin{cases}\dot{x} = Lx + C_2 Lt^2 \\ \dot{y} = \left(1 + \gamma Cm^3 t^2\right)y + 4x + \left(4\gamma^2 C_2 + 4LD + \gamma Cm^3\right)t^2 + 4\gamma^2 C^2 m^2 t^4\end{cases} \qquad \begin{cases}x(0) = 0 \\ y(0) = 0\end{cases} \tag{52}$$

Its solution can be analytically obtained, but omitted for the sake of length. ∎

H.3. Proof of Thm. 13

Proof [Proof of Lemma 12] By the definition of semi-distance ρ in Eq. (7), we have

$$\rho((\hat{g},\hat{\xi}),(\tilde{g},\tilde{\xi})) - \rho((\hat{g},\hat{\xi}),(g,\xi))$$
$$= f(r((\hat{g},\hat{\xi}),(\tilde{g},\tilde{\xi})))\left(1 + \beta\|\hat{\xi} - \tilde{\xi}\|^2\right)$$
$$- f(r((\hat{g},\hat{\xi}),(g,\xi)))\left(1 + \beta\|\hat{\xi} - \xi\|^2\right)$$
$$= \beta f(r((\hat{g},\hat{\xi}),(\tilde{g},\tilde{\xi})))\left(\|\hat{\xi} - \tilde{\xi}\|^2 - \|\hat{\xi} - \xi\|^2\right)$$
$$+ \left(f(r((\hat{g},\hat{\xi}),(\tilde{g},\tilde{\xi}))) - f(r((\hat{g},\hat{\xi}),(g,\xi)))\left(1 + \beta\|\hat{\xi} - \xi\|^2\right)\right)$$

We will bound the 2 terms in the last equation separately.

The first term:

$$
\begin{aligned}
\left\|\hat{\xi} - \tilde{\xi}\right\|^2 &- \left\|\hat{\xi} - \xi\right\|^2 \\
&= \left\|\tilde{\xi} - \xi\right\| \left\|2\hat{\xi} - \tilde{\xi} - \xi\right\| \\
&\le 2\delta \left\|\hat{\xi} - \xi\right\| + \delta^2 \\
&\le \left\|\hat{\xi} - \xi\right\|^2 + 2\delta^2
\end{aligned}
$$

and as a result,

$$
\begin{aligned}
\beta f(r((\hat{g}, \hat{\xi}), (g, \xi))) &\left(\left\|\hat{\xi} - \tilde{\xi}\right\|^2 - \left\|\hat{\xi} - \xi\right\|^2 \right) \\
&\le \beta f(r((\hat{g}, \hat{\xi}), (g, \xi))) \left(\left\|\hat{\xi} - \xi\right\|^2 + 2\delta^2 \right) \\
&\le \rho((\hat{g}, \hat{\xi}), (g, \xi))\delta + 2\beta f(R)\delta^2
\end{aligned}
$$

The second term: Because $f' \le 1$

$$
\begin{aligned}
(f(r((\hat{g}, \hat{\xi}), (\tilde{g}, \tilde{\xi}))) &- f(r((\hat{g}, \hat{\xi}), (g, \xi)))) \left(1 + \beta\left\|\hat{\xi} - \xi\right\|^2\right) \\
&\le (r((\hat{g}, \hat{\xi}), (\tilde{g}, \tilde{\xi}))) - r((\hat{g}, \hat{\xi}), (g, \xi))) \left(1 + \beta\left\|\hat{\xi} - \xi\right\|^2\right) \\
&\le \left(1 + \alpha + \gamma^{-1}\right) \left(1 + \beta\left\|\hat{\xi} - \xi\right\|^2\right)\delta \\
&\le \left(1 + \alpha + \gamma^{-1}\right) \frac{f(r((\hat{g}, \hat{\xi}), (g, \xi))))}{f(R)} \left(1 + \beta\left\|\hat{\xi} - \xi\right\|^2\right)\delta \\
&= \left(1 + \alpha + \gamma^{-1}\right) \frac{1}{f(R)} \rho((\hat{g}, \hat{\xi}), (g, \xi))\delta
\end{aligned}
$$

Sum them up we have

$$
\begin{aligned}
\rho((\hat{g}, \hat{\xi}), (\tilde{g}, \tilde{\xi})) &- \rho((\hat{g}, \hat{\xi}), (g, \xi)) \\
&\le \left(1 + \frac{1 + \alpha + \gamma^{-1}}{f(R)}\right) \rho((\hat{g}, \hat{\xi}), (g, \xi))\delta + 2\beta f(R)\delta^2
\end{aligned}
$$

We have the desired result by defining

$$
\begin{cases}
A_1 := 1 + \frac{1+\alpha+\gamma^{-1}}{f(R)} \\
A_2 := 2\beta f(R)
\end{cases}
\tag{53}
$$

$\blacksquare$

Proof [Proof of Lemma 11] We focus on x first. Since $\dot{x} \le Lx + C_2 L t_0^2$, we have $x(t) \le C_2 t_0^2 (e^{Lt} - 1) = \mathcal{O}(t^3)$.

For y, we have

$$
\begin{aligned}
\dot{y} &\le \left(1 + \gamma C m^3 t_0^2\right) y + 4x + \left(4\gamma^2 C_2 + 4LD + \gamma C m^3\right) t_0^2 + 4\gamma^2 C^2 m^2 t_0^4 \\
&\le \left(1 + \gamma C m^3 t_0^2\right) y + 4C_2 t_0^2 e^{Lt} + \left(4\gamma^2 C_2 - 4C_2 + 4LD + \gamma C m^3\right) t_0^2 + 4\gamma^2 C^2 m^2 t_0^4
\end{aligned}
$$

Solving the ODE explicitly, we have

$$y \le \frac{4\gamma^2 C_2 - 4C_2 + 4LD + \gamma Cm^3 + 4\gamma^2 C^2 m^2 t_0^2}{1 + \gamma Cm^3 t_0^2} t_0^2 \left(e^{(1+\gamma Cm^3 t_0^2)t} - 1 \right) + \frac{e^{(1+\gamma Cm^3 t_0^2)t} - e^{Lt}}{1 + \gamma Cm^3 t_0^2 - L} 4C_2 t_0^2$$

Since we have $e^t - 1 = \mathcal{O}(t)$, we have $y(t) = \mathcal{O}(t^3)$. ∎

Corollary 47 *Suppose the Lie group* G *is compact with diameter* D. *For any* (g, ξ) *and* $(\hat{g}, \hat{\xi})$ *in the following cases:*

- *(g, ξ) is the solution of sampling SDE Eq.* (4) *at time* t *for any* $t \ge 0$.
  $(\hat{g}, \hat{\xi})$ is the solution of sampling SDE Eq. (4) *at time* s *for any* $s \ge 0$.

- *(g, ξ) is the solution of numerical discretization Eq.* (14) *at step* k *for any* $k \in \mathbb{N}$.
  $(\hat{g}, \hat{\xi})$ is the solution of numerical discretization Eq. (14) *at step* l *for any* $l \in \mathbb{N}$.

- *(g, ξ) is the solution of sampling SDE Eq.* (4) *at time* t *for any* $t \ge 0$.
  $(\hat{g}, \hat{\xi})$ is the solution of numerical discretization Eq. (14) *at step* k *for any* $k \in \mathbb{N}$.

then we have

$$\mathbb{E}\rho^2((\hat{g}, \hat{\xi}), (g, \xi)) \le f(R_1)^2 (1 + 8C_2 + 16C_4)$$

where C_2 *and* C_4 *are given in Lemma* 46.

Proof [Proof of Cor. 47] By our design that f is upper bounded by $f(R)$, we have

$$\rho((g, \xi), (\hat{g}, \hat{\xi})) \le f(R_1)(1 + \left\| \xi - \hat{\xi} \right\|^2)$$
$$\le f(R_1)(1 + 2\|\xi\|^2 + 2\|\hat{\xi}\|^2)$$

and by Cauchy-Schwarz inequality

$$\rho^2((g, \xi), (\hat{g}, \hat{\xi})) \le f(R_1)^2 \left(1 + 4\|\xi\|^2 + 4\|\hat{\xi}\|^2 + 8\|\xi\|^4 + 8\|\hat{\xi}\|^4 \right)$$

By taking expectation and Lemma 46, we have the corollary proved. ∎

Proof [Proof of Thm. 13] In this proof, we denote $\delta := d((\tilde{g}_t, \tilde{\xi}_t), (g_t, \xi_t))$. Using the modified triangle inequality (Lemma 12),

$$\mathbb{E}\rho((\hat{g}_t, \hat{\xi}_t), (\tilde{g}_t, \tilde{\xi}_t))$$
$$\le \mathbb{E}\left[(1 + A_1\delta)\rho((\hat{g}_t, \hat{\xi}_t), (g_t, \xi_t)) + A_2\delta^2 \right]$$
$$\le \exp(-ct)\mathbb{E}\rho((\hat{g}_0, \hat{\xi}_0), (g_0, \xi_0)) + A_1\mathbb{E}\delta\rho((\hat{g}_t, \hat{\xi}_t), (g_t, \xi_t)) + A_2\mathbb{E}\delta^2$$
$$\le \exp(-ct)\mathbb{E}\rho((\hat{g}_0, \hat{\xi}_0), (g_0, \xi_0)) + \frac{A_1}{2}\mathbb{E}\left[\frac{\delta^2}{t^{\frac{3}{2}}} + t^{\frac{3}{2}}\rho^2((\hat{g}_t, \hat{\xi}_t), (g_t, \xi_t)) \right] + A_2\mathbb{E}\delta^2$$

where the last step is because of the Cauchy-Schwarz inequality. By Cor. 47, we have $\mathbb{E}\rho$ and $\mathbb{E}\rho^2$ is bounded and

$$\mathbb{E}\rho((\hat{g}_t, \hat{\xi}_t), (\tilde{g}_t, \tilde{\xi}_t))$$
$$\le \exp(-ct)\mathbb{E}\rho((\hat{g}_0, \hat{\xi}_0), (g_0, \xi_0)) + \left(\frac{A_1}{2t^{\frac{3}{2}}} + A_2 \right)\mathbb{E}\delta^2 + \frac{A_1}{2}f(R)^2(1 + 8C_2 + 16C_4)t^{\frac{3}{2}}$$

∎

H.4. Proof of Cor. 14, Thm. 15 and Thm. 16

Proof [Proof of Cor. 14] We apply Thm. 13 recurrently, which gives

$$
\begin{aligned}
&\mathbb{E}\rho((\hat{g}_{kh}, \hat{\xi}_{kh}), (\tilde{g}_k, \tilde{\xi}_k)) \\
&\le e^{-ch}\mathbb{E}\rho((\hat{g}_{(k-1)h}, \hat{\xi}_{(k-1)h}), (\tilde{g}_{(k-1)h}, \tilde{\xi}_{(k-1)h})), (\tilde{g}_{k-1}, \tilde{\xi}_{k-1})) + E(h) \\
&\le e^{-2ch}\mathbb{E}\rho((\hat{g}_{(k-2)h}, \hat{\xi}_{(k-1)h}), (\tilde{g}_{(k-2)h}, \tilde{\xi}_{(k-2)h})), (\tilde{g}_{k-2}, \tilde{\xi}_{k-2})) + (1 + e^{-ch})E(h) \\
&\le \cdots \\
&\le e^{-ckh}\mathbb{E}\rho((\hat{g}_0, \hat{\xi}_0), (\tilde{g}_0, \tilde{\xi}_0)), (\tilde{g}_0, \tilde{\xi}_0)) + \sum_{i=0}^{k} e^{-cih}E(h) \\
&\le e^{-ckh}\mathbb{E}\rho((\hat{g}_0, \hat{\xi}_0), (\tilde{g}_0, \tilde{\xi}_0)), (\tilde{g}_0, \tilde{\xi}_0)) + \frac{E(h)}{1 - \exp(-ch)}
\end{aligned}
$$

■

Proof [Proof of Thm. 15] Let $\mathrm{Law}(\hat{g}_0, \hat{\xi}_0)$ follows the target distribution and $(\hat{g}_t, \hat{\xi}_t)$ is the solution of the sampling dynamics (4). As a result, $\mathrm{Law}(\hat{g}_t, \hat{\xi}_t) \sim \nu_*, \forall t$. We make $(\hat{g}_0, \hat{\xi}_0)$ and $(\tilde{g}_0, \tilde{\xi}_0)$ are paired in the optimal way under ρ, i.e., $\mathbb{E}\rho((\hat{g}_0, \hat{\xi}_0), (\tilde{g}_0, \tilde{\xi}_0)) = W_\rho(\mathrm{Law}(\hat{g}_0, \hat{\xi}_0), \mathrm{Law}(\tilde{g}_0, \tilde{\xi}_0))$.

Cor. 14 gives

$$
\begin{aligned}
&W_\rho(\tilde{\nu}_k, \nu_*) \\
&W_\rho(\mathrm{Law}(\hat{g}_k, \hat{\xi}_k), \mathrm{Law}(\tilde{g}_k, \tilde{\xi}_k)) \\
&\le \mathbb{E}\rho((\hat{g}_{kh}, \hat{\xi}_{kh}), (\tilde{g}_k, \tilde{\xi}_k)) \\
&\le e^{-ckh}\mathbb{E}\rho((\hat{g}_0, \hat{\xi}_0), (\tilde{g}_0, \tilde{\xi}_0)) + \frac{E(h)}{1 - \exp(-ch)} \\
&= e^{-ckh}W_\rho(\mathrm{Law}(\hat{g}_0, \hat{\xi}_0), \mathrm{Law}(\tilde{g}_0, \tilde{\xi}_0)) + \frac{E(h)}{1 - \exp(-ch)}
\end{aligned}
$$

■

Proof [Proof of Thm. 16] This is a direct corollary of Lemma 39 and Thm. 15. ■