Geodesic Slice Sampler for Multimodal Distributions with Strong Curvature

Bernardo Williams[1] **Hanlin Yu**[1] **Hoang Phuc Hau Luu**[1] **Georgios Arvanitidis**[2] **Arto Klami**[1]

[1]Department of Computer Science, University of Helsinki, Finland
[2]Cognitive Systems, DTU Compute, Technical University of Denmark

Abstract

Traditional Markov Chain Monte Carlo sampling methods often struggle with sharp curvatures, intricate geometries, and multimodal distributions. Slice sampling can resolve local exploration inefficiency issues, and Riemannian geometries help with sharp curvatures. Recent extensions enable slice sampling on Riemannian manifolds, but they are restricted to cases where geodesics are available in a closed form. We propose a method that generalizes Hit-and-Run slice sampling to more general geometries tailored to the target distribution, by approximating geodesics as solutions to differential equations. Our approach enables the exploration of the regions with strong curvature and rapid transitions between modes in multimodal distributions. We demonstrate the advantages of the approach over challenging sampling problems.

1 INTRODUCTION

Sampling from a differentiable unnormalized log-density defined on a Euclidean space is a core problem in machine learning and statistics. While gradient-based Markov Chain Monte Carlo (MCMC) methods have proven effective in many scenarios, they often face significant challenges when the target distribution exhibits complex geometry (sharp curvature) or multimodal behavior. The two core challenges are largely addressed with complementary techniques, with little work on algorithms that excel for targets that are *both* multimodal and complex in shape.

Complex shapes and sharp curvatures are often addressed by using a suitably chosen Riemannian geometry within the sampling algorithms [Girolami and Calderhead, 2011]. Instead of operating in a Euclidean space and metric, the samplers carry out the necessary operations using a metric that adapts to the curvature of the parameter space. In practice, the methods follow flows induced by the metric, in most cases by numerical integration, and consequently the methods are sometimes called *geodesic* methods as in our title. Various practical metrics and sampling algorithms have been shown to improve the sampling of targets with strong curvature [Girolami and Calderhead, 2011, Byrne and Girolami, 2013, Lan et al., 2015, Hartmann et al., 2022, 2023, Williams et al., 2024], albeit always with increased computational cost.

Multimodality, in turn, is most commonly addressed by tempering or diffusion techniques [Earl and Deem, 2005, Chen et al., 2024]. These methods use a tempered (smoothed) version of the target to improve exploration over multiple modes, intuitively changing the problem itself so that the modes are connected with areas of sufficient probability. At a high degree of tempering these methods can efficiently explore the different modes, but low tempering is needed for accurate sampling within the modes, necessitating adaptive or parallel sampling with different degrees of tempering. The efficiency of parallel tempering depends on the swap acceptance rate between adjacent temperatures, which can decrease in high dimensions if the temperature schedule is not well-tuned [Woodard et al., 2009]. Diffusion-based approaches, in turn, require careful choice of the noise schedule to balance exploration and accuracy [Song and Ermon, 2019, Chen et al., 2024]. Unlike tempering, diffusion methods can achieve smooth transitions between modes without explicitly maintaining a set of parallel chains, but the acceptance rate of noisy samples can be low [Chen et al., 2024].

Even though the two approaches are efficient in addressing the two challenges separately, there is very little work on samplers designed for the general setup where both difficulties may arise simultaneously. One could consider e.g. parallel tempering in a Riemannian metric — see Byrne and Girolami [2013] for a rare example in this intersection — but ideally we would like to address both aspects using a common mechanism. This work explores one such approach, developing a Riemannian sampler capable of efficiently exploring multiple modes, without any tempering

for the target distribution. Instead, we seek to improve mode exploration by changing the metric, in the spirit of the early work by Lan et al. [2014] that developed a specific metric solely for this purpose. Their metric, however, requires explicit identification and tracking of the modes and is more like a conceptual demonstration, and we are not aware of any other works aiming for efficient multimodal samplers solely by the change of the metric.

We are motivated by the idealized slice sampler with computable level sets. As noted by Durmus et al. [2023]: *"This means that the performance of the idealized slice sampler is ignorant of the introduction of, e.g., multimodality, local modes, or anisotropy as long as the volume of the level sets is not modified."* This insight suggests that by modifying the geometry of the problem to produce simpler or more tractable level sets, the slice sampler can effectively handle multimodal distributions. From a practical perspective, we build on the (Euclidean) Hit-and-Run slice sampler by Bélisle et al. [1993], which at each iteration selects a random direction and then samples from the resulting one-dimensional distribution formed by the intersection of the line and the slice. In effect, it transforms multi-dimensional sampling into sequential one-dimensional sampling tasks, but the overall sampler may be inefficient. Especially in higher dimensions, the intersection with the slice can be small for almost all directions Murray et al. [2010].

Both Habeck et al. [2023] and Durmus et al. [2023] recently considered generalizations of the Hit-and-Run sampler for Riemannian manifolds, replacing the lines with geodesics. We build on the general algorithmic framework introduced by Durmus et al. [2023] and adapt it to the task of sampling from a distribution with a complex geometry. Specifically, we begin by embedding the (Euclidean) sampling space into a higher-dimensional space that incorporates the target distribution's geometric information, such as Fisher information or Monge embedding [Hartmann et al., 2022]. This transforms the problem into sampling from a particular Riemannian manifold where the target distribution corresponds to the Hausdorff density (see Section 3). Note that even though we leverage components proposed by Durmus et al. [2023], our task is fundamentally more difficult. Their starting point was sampling of a density on a known manifold (e.g., a sphere) where the geodesics are exactly known, whereas the complexity of our embedding manifold requires us to approximate the geodesics using numerical integrators.

In this work, we propose a geodesic slice sampler applicable for arbitrary Riemannian metrics, and discuss the choice of the metric. In particular, we introduce two new computationally efficient metrics. Both metrics improve sampling over multimodal targets by, in a sense, pulling the modes closer to each other; see Figure 1 illustrating this effect within the slice sampler, as a function of a parameter λ controlling how much the metric warps the space. In addition, we introduce a meta-sampler similar to Tjelmeland and Hegstad [2001]

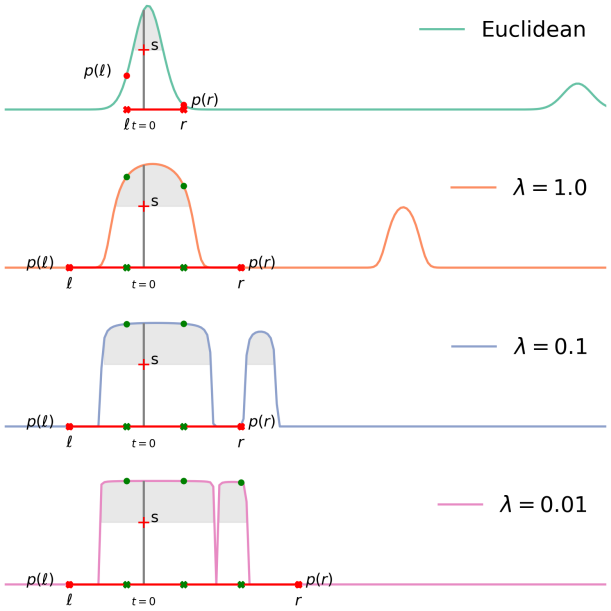

Figure 1: Illustration of the step-out procedure in Metric-agnostic Geodesic Slice Sampler. The lines drawn with different colors represent the Hausdorff density $p(t) := p_{\mathcal{H}}(\hat{\gamma}_{(\boldsymbol{x},\boldsymbol{v})}(t))$ (Eq. (1)) considering the Inverse Generative metric for different values of λ (Eq. (4)). The step-out procedure chooses $s \sim \mathrm{Unif}(0, p(0))$ and sets randomly an interval of length $r - \ell$ at $t = 0$ with left and right points ℓ and r. While $p(r) > p(s)$ it expands the right side of the interval as $r = r + w$ and for the left side while $p(l) > p(s)$ it does $\ell = \ell - w$. This expands the length of the initial interval. As $\lambda \to 0$ the space shrinks due the properties of the metric, making it easier for the step-out procedure to jump to the distant mode.

that combines the proposed method with a separate sampler for improved exploration of local modes.

We empirically demonstrate improved sampling over Euclidean methods for complex targets, and highlight improved mixing over multiple modes in high dimensional-cases when compared against parallel tempering [Swendsen and Wang, 1986, Łatuszyński et al., 2025] and the diffusive Gibbs sampler by Chen et al. [2024] designed for addressing multimodality. Similar to previous Riemannian methods, the algorithm shows good exploration and mixing, but has slower iterations because of the numerical computation of the geodesics.

2 BACKGROUND: SLICE SAMPLING

The classic work of Neal [2003] introduces slice sampling as a method for generating samples by uniformly sampling from the $\mathbb{R}^{D+1}$ manifold defined by the graph of the probability density. Let $p(\boldsymbol{x})$ be an unnormalized continuous

target density that satisfies $\int p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} < \infty$. Suppose that direct sampling from $p(\boldsymbol{x})$ is not feasible. We consider densities where $\boldsymbol{x} \in \mathbb{R}^D$ with respect to the Lebesgue measure.

Idealized slice sampling defines a uniform distribution over the volume under the graph of $p(\boldsymbol{x})$ and generates samples through the following two steps:

1. Sample $s \sim \mathrm{Unif}(0, p(\boldsymbol{x}))$.
2. Sample $\boldsymbol{x} \sim \mathrm{Unif}(L(s))$.

where the slice is given by $L(s) := \{\boldsymbol{x} \mid p(\boldsymbol{x}) > s\}$. For special cases, such as log-concave or rotationally invariant densities, the slice sampler has theoretical performance guarantees [Natarovskii et al., 2021]. However, for more complex distributions, drawing uniform samples from $L(s)$ is often impractical [Rudolf and Ullrich, 2018].

To address this, the step-out and shrinkage procedures are used. Below, we provide an informal explanation of these procedures. The full algorithm is detailed in the Appendix (Algorithms 3 and 4). Both procedures were first introduced by Neal [2003], but we adopt an equally valid modified version of the shrinkage step as proposed by Durmus et al. [2023]. For a moment, assume a univariate density $p(x)$ and a current position $x \in \mathbb{R}$. The procedures are as follows:

The Step-Out Procedure The step-out procedure, illustrated in Figure 1, takes two parameters: the width $w \in \mathbb{R}$ and maximum steps $m \in \mathbb{N}$. Given the slice $L(s)$, the goal is to expand an interval around the current point x. Consider the auxiliary function $\gamma_x(t) = x + t$.

The initial left ℓ and right r points are set at a random distance w apart. This is done by sampling $u \sim \mathrm{Unif}(0, w)$ and setting $\ell = -u$ and $r = \ell + w$. To ensure that at most $m + 1$ expansion steps are performed (combined for both directions), a random integer $\iota \sim \mathrm{Unif}(\{1, \ldots, m\})$ is sampled. The right limit is expanded up to ι times, and the left limit up to $m + 1 - \iota$ times.

The expansion proceeds as follows: The right limit r is expanded by adding w until $p(\gamma_x(r + w)) < s$, meaning $\gamma_x(r + w) \notin L(s)$. The left limit ℓ is expanded by subtracting w until $p(\gamma_x(\ell - w)) < s$, meaning $\gamma_x(\ell - w) \notin L(s)$. The procedure returns the updated interval (ℓ, r). We denote it by Step-out$_{w,m}(s, \gamma_x)$.

The Shrinkage Procedure The shrinkage procedure selects a sample from the interval (ℓ, r) by gradually reducing its size until a point is found within $L(s) \cap (\ell, r)$.

The interval $J = (\ell, r)$ is treated as a circular domain, meaning that if we move past r, we continue from ℓ. The procedure starts by sampling two points y and z uniformly within (ℓ, r). If neither $\gamma_x(y)$ or $\gamma_x(z)$ fall inside $L(s)$, the interval is shrunk as follows:

- Form the interval $(y \wedge z, y \vee z)$. Update the circular region by

$$J = \begin{cases} J \cap (y \wedge z, y \vee z), & \text{if } 0 \in J, \\ J \setminus (y \wedge z, y \vee z), & \text{if } 0 \notin J. \end{cases}$$

- Set $y = z$ and update $z \sim \mathrm{Unif}(J)$.
- This process repeats, each time reducing the size of the interval, until $\gamma_x(z) \in L(s)$.

We denote the procedure Shrink$_{l,r}(s, \gamma_x)$. One complete step of the slice sampler is:

1. Sample $s \sim \mathrm{Unif}(0, p(x))$
2. Obtain $\ell, r =$ Step-out$_{w,m}(s, \gamma_x)$
3. Sample $t^* =$ Shrink$_{\ell,r}(s, \gamma_x)$.
4. Set $x = \gamma_x(t^*)$.

Hit-and-Run One way to extend slice sampling to multivariate distributions is to combine it with Hit-and-Run sampling, presented here following Bélisle et al. [1993]. Let $\mathbb{S}^{D-1}(\boldsymbol{x}) = \{\boldsymbol{v} \in \mathbb{R}^D : \|\boldsymbol{v}\|^2 = 1\}$, and $\boldsymbol{v} \sim \mathrm{Unif}(\mathbb{S}^{D-1}(\boldsymbol{x}))$. An iteration of the whole sampler is:

- Obtain $\boldsymbol{v} \sim \mathrm{Unif}(\mathbb{S}^{D-1}(\boldsymbol{x}))$.
- Obtain a sample from the density evaluated on the straight line (Euclidean geodesic)

$$t \mapsto p(\boldsymbol{x} + t\,\boldsymbol{v}) / \int p(\boldsymbol{x} + t\,\boldsymbol{v})\mathrm{d}t.$$

When directly sampling a value t according the density along a straight line $t \mapsto p(\boldsymbol{x} + t\,\boldsymbol{v}) / \int p(\boldsymbol{x} + t\,\boldsymbol{v})\mathrm{d}t$ is not feasible, we can use slice sampling on $t \mapsto p(\boldsymbol{x} + t\,\boldsymbol{v})$, since it is an unnormalized univariate distribution. Define $\gamma_{(x,v)}(t) = \boldsymbol{x} + t\,\boldsymbol{v}$. The step-out procedure outputs $\ell, r =$ Step-out$_{w,m}(s, \gamma_{(x,v)})$ and the shrinkage procedure will return $t^* =$ Shrink$_{\ell,r}(s, \gamma_{(x,v)})$. The new sample is $\boldsymbol{x} = \gamma_{(x,v)}(t^*)$. This is called Hit-and-Run slice sampling or hybrid slice sampling [Łatuszyński and Rudolf, 2014]. This method extends slice sampling to probability distributions defined over $\mathbb{R}^D$.

3 METHOD

Our main contribution is a geodesic slice sampler that can accommodate arbitrary metrics. It extends the Hit-and-Run slice sampler described above for non-Euclidean geometries, similar to the recent works of Durmus et al. [2023] and Habeck et al. [2023], but instead of leveraging closed-form analytic geodesics of predefined manifolds we induce metrics using characteristics of the target density itself to guide the sampling. Now the geodesics need to be approximated by numerical integrators. This section explains the sampler and a meta-sampler that combines the core method with

separate local sampler for improved efficiency for general metrics, always using $\boldsymbol{G}(\boldsymbol{x})$ to denote the metric tensor. We will discuss specific metrics in Section 4.

Straight-line Hit-and-Run sampling can be inefficient because proposals often move away from high-probability regions [Murray et al., 2010]. To resolve this, we perform slice sampling along geodesic curves that can accommodate the geometry of the target distribution. This improves efficiency when the target distribution is highly curved or multimodal; see Section 4.2. We are interested in sampling problems defined in $\mathbb{R}^D$ but allow using different plug-in metrics (preferably using the target density information) to enhance exploration.

The general problem can be cast as sampling from a distribution defined on a Riemannian manifold where we adapt the algorithm of Durmus et al. [2023] under general metrics. Alternatively, it can be seen as Hit-and-Run where straight lines are replaced by curves that better wrap around the level sets of the target density (given the metrics are good enough). Because the metric is general, closed-form geodesics are unavailable, so we must compute them with numerical integrators. To correctly sample along geodesics, we need three key components: Adjusting for the correct density on the manifold, properly sampling directions using the Riemannian metric, and solving the geodesic equations.

Hausdorff Density: To ensure we sample from the correct distribution on the manifold with metric $\boldsymbol{G}(\boldsymbol{x})$, we must account for the change in measure from the Euclidean space to the manifold. The correct density is the Hausdorff density

$$p_{\mathcal{H}}(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{\sqrt{\det \boldsymbol{G}(\boldsymbol{x})}}. \tag{1}$$

The denominator adjusts for local volume distortion introduced by the metric $\boldsymbol{G}(\boldsymbol{x})$, ensuring that the volume over the manifold is preserved and hence maintaining proper sampling behavior. See Appendix A.5 for further details.

Sampling from the Riemannian Unit Ball: Instead of sampling a random direction in Euclidean space, we must now sample from the unit geodesic ball under the Riemannian metric, where we can directly use the method proposed by Durmus et al. [2023]. Given a position $\boldsymbol{x}$, a velocity $\boldsymbol{v}$ is sampled as follows: First draw $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}^{-1}(\boldsymbol{x}))$, and then normalize it to obtain a unit-length vector in the Riemannian metric with:

$$\boldsymbol{v} \leftarrow \frac{\boldsymbol{v}}{\| \boldsymbol{v} \|_g}, \quad \text{where} \quad \| \boldsymbol{v} \|_g = \sqrt{\boldsymbol{v}^\top \boldsymbol{G}(\boldsymbol{x})\, \boldsymbol{v}}.$$

This ensures that the direction is uniformly distributed on the unit sphere under the metric $\boldsymbol{G}(\boldsymbol{x})$. See Appendix A for additional implementation details.

Approximating Geodesic Curves Given a sampled velocity $\boldsymbol{v}$, we need to follow the geodesic curve starting at $\boldsymbol{x}$ in direction $\boldsymbol{v}$. In general, the geodesic equation

$$\begin{aligned} \dot{\boldsymbol{x}}_k &= \boldsymbol{v}_k, \\ \dot{\boldsymbol{v}}_k &= -\|\boldsymbol{v}\|^2_{\Gamma^k}, \quad \text{for } k = 1, \ldots, D. \end{aligned} \tag{2}$$

where $\Gamma^k_{ij} = \frac{1}{2} g^{km}(\partial_i g_{mj} + \partial_j g_{im} - \partial_m g_{ij})$, does not have a closed-form solution for arbitrary $\boldsymbol{G}(\boldsymbol{x})$. See more detail in appendix A.3. Instead, we numerically approximate the exponential map $\gamma_{(\boldsymbol{x},\boldsymbol{v})}(t)$ by solving these differential equations with an ordinary differential equation (ODE) solver, denoted as $\hat{\gamma}_{(\boldsymbol{x},\boldsymbol{v})}(t)$. The choice of the metric determines the shape of geodesic trajectories, allowing the sampler to adapt to different target distributions; see Section 4.

Algorithm 1 explains the full Metric-agnostic Geodesic Slice Sampler (MAGSS). After sampling a velocity $\boldsymbol{v}$ from the unit Riemannian sphere, slice sampling is performed on the Hausdorff density evaluated along the numerical solution of the geodesic trajectory. The step-out and shrinkage procedures then determine the final sample.

Algorithm 1 Metric-agnostic Geodesic Slice Sampler

Input: Initial position $\boldsymbol{x}^{[0]}$, metric tensor $\boldsymbol{G}(\boldsymbol{x})$, and parameters $m \in \mathbb{N}$, $w \geq 0$.
Output: N samples $\boldsymbol{x}^{[n]}$.

1: **for** $n \leftarrow 0, \ldots, N-1$ **do**
2: Sample $s \sim \mathrm{Unif}(0, p_{\mathcal{H}}(\boldsymbol{x}^{[n]}))$
3: Sample velocity $\boldsymbol{v}^{[n]} \sim \mathrm{Unif}(\mathbb{S}^{D-1}_g(\boldsymbol{x}^{[n]}))$
4: Compute $(\ell, r) = \text{Step-out}_{w,m}(s, \hat{\gamma}_{(\boldsymbol{x}^{[n]}, \boldsymbol{v}^{[n]})})$
5: Sample time $t^* = \text{Shrink}_{\ell,r}(s, \hat{\gamma}_{(\boldsymbol{x}^{[n]}, \boldsymbol{v}^{[n]})})$
6: $\boldsymbol{x}^{[n+1]} = \hat{\gamma}_{(\boldsymbol{x}^{[n]}, \boldsymbol{v}^{[n]})}(t^*)$
7: **end for**

3.1 META SAMPLER AND MULTIMODALITY

The sampler as described above is valid as such, but we also introduce a simple extension that can further improve sampling for multimodal targets with complex local structure.

Following Tjelmeland and Hegstad [2001], Łatuszyński et al. [2025], we create a *meta-sampler* that alternates between using MAGSS for global moves and an arbitrary local MCMC for sampling within each mode. To generate one sample, we first run K-steps of MAGSS followed by L-steps of any local MCMC sampler. We refer to this combined strategy as Meta-MAGSS, detailed in Algorithm 2 (in Appendix). The main motivation for this hybrid strategy is to leverage gradient-based algorithms for fast exploration of the mode, to utilize their efficient mixing and fast per-iteration computation when they are sufficiently good for the local target. We could in principle use any sampler for the local part, including Riemannian samplers, but we in practice use standard Euclidean Metropolis-adjusted Langevin Algorithms (MALA) [Roberts and Tweedie, 1996] in our experiments.

4 METRICS

The sampler is general, applicable for an arbitrary metric and only requiring $\boldsymbol{G}(\boldsymbol{x})$ to be positive definite and vary continuously. By selecting an appropriate metric we can influence how the geodesics explore the space, controlling the overall sampling behavior. There is no single metric that is optimal for all targets, and the metrics proposed in the literature are motivated by complementary argumentation, with notable emphasis in computational efficiency.

Next we discuss the metric choice. The literature has exclusively focused on metrics that improve local exploration for complex target distributions, with several practical solutions that we re-cap in Section 4.1. We then turn our attention on how to improve exploration of multiple modes, presenting novel metrics specifically designed for this in Section 4.2.

4.1 FOR ADAPTING TO LOCAL CURVATURE

The Fisher metric The Fisher Information Metric (FIM) is defined as the covariance of the score function, and was predominantly used in the early Riemannian methods [Girolami and Calderhead, 2011] due to its close connection to estimation theory. A general form of the metric is:

$$\boldsymbol{G}_F(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{y}\,|\,\boldsymbol{x}}\left[\nabla_{\boldsymbol{x}} \log p(\boldsymbol{y}\,|\,\boldsymbol{x}) \nabla_{\boldsymbol{x}} \log p(\boldsymbol{y}\,|\,\boldsymbol{x})^\top\right],$$

but the specific form depends on the underlying problem, due to integration over the conditional density. Furthermore, it requires direct matrix inversion for computing $\boldsymbol{G}_F^{-1}(\boldsymbol{x})$ that is required during geodesic computations (Eq. 2), with complexity of $\mathcal{O}(D^3)$. This makes the metric impractical and inefficient for high-dimensional problems.

The Monge Metric The computational cost of solving the geodesic equations (Eq. 2) is primarily determined by the inversion of the metric tensor, and consequently metrics with closed-form inverse offer significant savings. The Monge metric by Hartmann et al. [2022] naturally arises from the geometry of the graph of log-density function when viewed as a submanifold embedded in $\mathbb{R}^{D+1}$. Let $\alpha^2 \geq 0$ and $\lambda \geq 0$. The Monge metric and its inverse are given by

$$\begin{aligned}\boldsymbol{G}_M(\boldsymbol{x}) &= \boldsymbol{I}_D + \alpha^2 \nabla\ell \nabla\ell^\top, \\ \boldsymbol{G}_M^{-1}(\boldsymbol{x}) &= \boldsymbol{I}_D - \frac{\alpha^2}{1+\alpha^2\|\nabla\ell\|^2} \nabla\ell\nabla\ell^\top,\end{aligned} \tag{3}$$

where $\ell(x) = \ln p(x)$. As $\alpha^2 \to 0$, the metric reduces to the Euclidean metric $\boldsymbol{I}_D$. The determinant required for computing the Hausdorff density (Eq. (1)) is $\det \boldsymbol{G}_M(\boldsymbol{x}) = 1 + \alpha^2\|\nabla\ell\|^2$. Figure 2 illustrates the exponential map of geodesic balls with increasing radius under the Monge metric. This metric adapts to the geometry of the target distribution, expanding regions based on the local structure of the density.

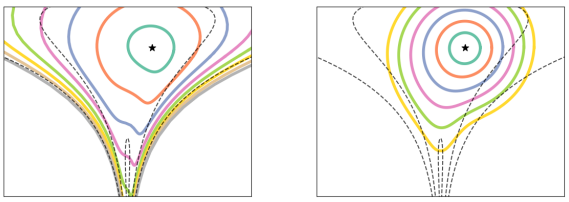

Figure 2: Exponential map for Riemannian balls of increasing radius on the Funnel distribution for the Monge metric with $\alpha = 1$ on the left panel. On the right, the plot is analogous but considering the Generative metric with $\lambda = 0.1$ and $p_0 = 0.1$. Each color represents a bigger radius from the base point ($\star$). Both metrics achieve the desired goal, shortening the distances to the points along the narrow funnel that would be difficult to reach in a Euclidean geometry.

The Generative Metric Another efficient metric, the Generative metric that is proportional to the target density function, was recently proposed by Kim et al. [2024]. One of its advantages is that computing the Christoffel symbols Γ_{ij}^k only requires first-order derivatives of the density, whereas the Monge metric (Equation 3) introduces second-order terms. For scalars $p_0 > 0$ and $\lambda \geq 0$, the Generative metric and its inverse are:

$$\boldsymbol{G}_g(\boldsymbol{x}) = \left(\frac{p_0 + \lambda}{p(\boldsymbol{x}) + \lambda}\right)^2 \boldsymbol{I}_D, \tag{4}$$

$$\boldsymbol{G}_g^{-1}(\boldsymbol{x}) = \left(\frac{p(\boldsymbol{x}) + \lambda}{p_0 + \lambda}\right)^2 \boldsymbol{I}_D. \tag{5}$$

As $\lambda \to \infty$, the metric reduces to the Euclidean metric. The determinant is $\det \boldsymbol{G}_g(\boldsymbol{x}) = \left(\frac{p_0+\lambda}{p(\boldsymbol{x})+\lambda}\right)^{2D}$. Figure 1 illustrates the effect of λ on the Hausdorff density along geodesics $t \mapsto p_{\mathcal{H}}\left(\hat{\gamma}_{(\boldsymbol{x},\boldsymbol{v})}(t)\right)$, and Figure 2 again shows how the Generative metric transforms the space.

4.2 FOR BRIDGING THE MODES

The above metrics adapt for the local curvature and have been designed to improve sampling of, for instance, narrow funnels by re-defining the proximity (see Figure 2). For assisting exploration of multimodal targets we need different kinds of metrics: Now we would want a metric that makes modes that are far away in the original Euclidean sense appear closer. With the exception of the construction of Lan et al. [2014], which we will discuss in Section 6, we are not aware of any previous metrics designed for this. Next, we introduce two computationally efficient metrics, with fast inverses and determinants, for assisting multimodal sampling.

Geodesic curves maintain a constant velocity norm in the Riemannian sense by construction. Let $\boldsymbol{x}_t = \gamma_{(\boldsymbol{x}_0,\boldsymbol{v}_0)}(t)$ be a geodesic curve with velocity $\boldsymbol{v}_t = \dot{\gamma}_{(\boldsymbol{x}_0,\boldsymbol{v}_0)}(t)$, starting from $\boldsymbol{x}_0$ with initial velocity $\boldsymbol{v}_0$. If the geodesic moves

toward a low-probability region where $p(\boldsymbol{x}_t) \to 0$, then the "mode bridging" behavior occurs if the Euclidean velocity norms satisfy $\|\boldsymbol{v}_0\|_2 \ll \|\boldsymbol{v}_t\|_2$. This means that as t increases in low-density regions, the geodesics curves accelerate and locally pull the distant modes closer.

We propose two metrics with the desired behavior, by leveraging the metrics described in Section 4.1 in a novel way. Any matrix $\boldsymbol{G}(\boldsymbol{x})$ defines a valid Riemannian metric as long as it is positive definite for every $\boldsymbol{x} \in \mathcal{M}$ and varies continuously on $\mathcal{M}$. Since the inverse of a positive definite matrix is also positive definite, we observe that it is possible to use any of the previously formulated $\boldsymbol{G}^{-1}(\boldsymbol{x})$ for defining a metric. This gives two new metrics that both help exploring multiple modes in different ways:

The Inverse Monge Metric We use

$$\boldsymbol{G}_{IM}(\boldsymbol{x}) = \boldsymbol{I}_D - \frac{\alpha^2}{1 + \alpha^2 \|\nabla \ell\|^2} \nabla \ell \nabla \ell^\top,$$

as the metric tensor, with the inverse $\boldsymbol{G}_{IM}^{-1}(\boldsymbol{x}) = \boldsymbol{G}_M(\boldsymbol{x})$ given by the previously introduced metric tensor of the standard Monge metric (Eq. (3)). The determinant of this metric is $1/\det(\boldsymbol{G}_M)$, and hence it retains the computational efficiency of the original Monge metric. Figure 3 illustrates the geodesics emanating from one mode of a bimodal distribution under the Inverse Monge metric. The metric twists the curves towards the second mode and slightly increases acceleration (seen by the change of color). Observation 1 mathematically states the conditions for the change of acceleration caused by the metric.

Observation 1. Let $p(\boldsymbol{x})$ be a smooth density function. Let $(\boldsymbol{x}_t, \boldsymbol{v}_t)$ be the geodesic flow with initial conditions $(\boldsymbol{x}_0, \boldsymbol{v}_0)$ with respect to the Inverse Monge metric, such that $\boldsymbol{x}_0$ is a local maximum. Then $\|\boldsymbol{v}_t\|_2 \geq \|\boldsymbol{v}_0\|_2$ for all $t \neq 0$.

The Inverse Generative Metric We use

$$\boldsymbol{G}_{Ig}(\boldsymbol{x}) = \left(\frac{p(\boldsymbol{x}) + \lambda}{p_0 + \lambda} \right)^2 \boldsymbol{I}_D,$$

and obtain the inverse $\boldsymbol{G}_{Ig}^{-1}(\boldsymbol{x}) = \boldsymbol{G}_g(\boldsymbol{x})$ as the metric tensor of the standard Generative metric (Eq. (4)) and the determinant as $1/\det(\boldsymbol{G}_g)$. Again, the computational efficiency of the original Generative metric is retained. Figure 3 illustrates the main effect of the metric, that is to accelerate on low density regions (indicated by the light color); it also twists the trajectories slightly (best seen within the initial mode and beyond the second mode in the top right corner). Additionally Figure 1 illustrates the behavior in a univariate distribution. The acceleration behavior is mathematically stated in Observation 2.

Observation 2. Let $p(\boldsymbol{x})$ be a smooth density function. Let $(\boldsymbol{x}_t, \boldsymbol{v}_t)$ be the geodesic flow with initial conditions $(\boldsymbol{x}_0, \boldsymbol{v}_0)$ such that $p(x_0) > 0$ with respect to the Inverse Generative metric. Then, for t such that $p(\boldsymbol{x}_t) \to 0$ we have $\|\boldsymbol{v}_t\|_2 > \|\boldsymbol{v}_t\|_0$.

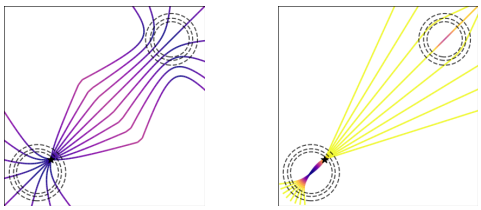

Figure 3: Effect of the metric for multimodal targets, showing the geodesics (lines) and the relative compression of the distance (color; yellow means faster travel in that area, darker colors mean slower travel). **Left:** Inverse Monge metric ($\alpha = 0.001$) helps more the geodesics to reach the other mode, and also slightly compresses the distances in the low-probability region. **Right:** Inverse Generative metric ($\lambda = 1$) compresses the distances in the low-probability region, but twists the paths only slightly.

The mathematical details for Observations 1 and 2 are given in Appendix A.6.

5 EXPERIMENTS

We evaluate MAGSS for targets with sharp curvature (Section 5.1), multiple modes (Section 5.2), or both (Section 5.3), always considering different choices of the metric. We also empirically quantify the effect of the numerical integrator. A code reproducing the experiments is available at github.com/williwilliams3/magss.

Evaluation We use primarily targets with known reference samples, which allows measuring the accuracy using the 1-Wasserstein (earth mover's) distance with the samples provided by the algorithm [Flamary et al., 2021]. Besides accuracy, we quantify the samplers with the probability of jumping between the different modes, as the raw ratio of consecutive samples that are within separate modes (defined manually for each problem).

Comparison methods To showcase the effect of the metric we will be running MAGSS also with in Euclidean metric, with $\boldsymbol{G}(\boldsymbol{x}) = \boldsymbol{I}_D$, and we additionally compared against the No-U-Turn Sampler, parallel tempering and diffusive Gibbs sampling.

The No-U-Turn Sampler (NUTS) is an auxiliary-variable sampler that augments the position $\boldsymbol{x}_t$ with a velocity $\boldsymbol{v}_t$ which jointly follow the Hamiltonian dynamics [Neal et al., 2011]. It adaptively determines the integration time by stopping at the first U-turn, i.e., the first time $t > 0$ such that $\langle \boldsymbol{x}_t - \boldsymbol{x}_0, \boldsymbol{v}_t \rangle < 0$ [Hoffman et al., 2014].

Parallel tempering (PT) runs many Markov Chains in parallel, each of which has $p(\boldsymbol{x})^{1/\tau_i}$ as targets for different temperatures $\tau_i \geq 1$, with $\tau_1 = 1$ recovering the original target. As $\tau \to \infty$, the density flattens, facilitating transi-

tions between regions of higher densities that are far apart from each other. The parallel chains jumps randomly between each other, thus visiting the modes more often according to a Metropolis-Hastings ratio [Swendsen and Wang, 1986, Geyer, 1991]. Our implementation of PT follows Łatuszyński et al. [2025].

Diffusive Gibbs sampling (DiGS) by Chen et al. [2024] is a sampler designed for addressing multimodality. It approaches the sampling task by using an auxiliary variable $\tilde{\boldsymbol{x}}$ with a Gibbs scheme. It uses the variance preserving (VP) [Song et al., 2021] noise scaling: $p(\tilde{\boldsymbol{x}}|\,\boldsymbol{x}) = \mathcal{N}(\tilde{\boldsymbol{x}}|\alpha_t\,\boldsymbol{x}, \sigma_t^2)$, where $\sigma_t = \sqrt{1-\alpha_t^2}$, sampled directly and $p(\boldsymbol{x}\,|\tilde{\boldsymbol{x}}) \propto p(\tilde{\boldsymbol{x}}|\,\boldsymbol{x})p(\boldsymbol{x})$ sampled through a local MCMC sampler. It has an additional Metropolis within Gibbs proposal scheme $q(\boldsymbol{x}\,|\tilde{\boldsymbol{x}}) = \mathcal{N}(\boldsymbol{x}\,|\tilde{\boldsymbol{x}}/\alpha_t, (\alpha_t/\sigma_t)^2)$. VP has the property that at when $\alpha_t \to 0$ then $p(\tilde{\boldsymbol{x}}|\,\boldsymbol{x}) = \mathcal{N}(\tilde{\boldsymbol{x}}|0, \boldsymbol{I}_D)$ and when $\alpha_t \to 1$ then, informally, $p(\tilde{\boldsymbol{x}}|\,\boldsymbol{x}) = \delta_{\boldsymbol{x}}$.

5.1 COMPLEX UNIMODAL TARGETS

We evaluate the methods on three canonical benchmark targets (funnel, hybrid Rosenbrock and squiggle) which exhibit strong curvature. The densities are given in Appendix C.5. Since these targets are unimodal, we only consider the metrics presented in Section 4.1 and exclude PT.

Figure 4 shows that MAGSS with Fisher metric $\boldsymbol{G}_F(\boldsymbol{x})$ is clearly superior, but runs out of the limited computational budget already at low dimensions, and the Monge metric $\boldsymbol{G}_M(\boldsymbol{x})$ offers notable improvement for Rosenbrock and squiggle targets. DiGS remains on the level of the Euclidean MAGSS and the Generative metric does not help either.

Experiment specification: We obtain $10,000$ samples using 10 chains and omit results for runs that did not complete in 12 hours. We set $\alpha^2 = 1$ for the Monge metric since this value has been shown to work [Hartmann et al., 2022]. We select $\lambda = 1$, $p_0 = 1$ for the Generative metric without further tuning. We use Dopri5 integrator with adaptive step-size. We set $w = 3$ and $m = 8$. DiGS and NUTS uses a single noise scale $\alpha = 1$ and step-size 0.1 for MALA within the algorithm.

5.2 MULTIMODAL WITH SIMPLE MODES

For studying mode exploration, we use a target of two D-dimensional Gaussian distributions centered at $-\mathbf{1}_D$ and $\mathbf{1}_D$ with scale $\sigma = 0.1$ and weights $\{0.2, 0.8\}$. The distance between the modes ($\sqrt{D}2$) grows for increasing dimensions, making transition between the modes more difficult. Now we only consider the new metrics for boosting mixing between the modes (Section 4.2).

Figure 5 reports the corresponding accuracies and reports the percentage of jumps between the modes. While the comparison methods PT and DiGS explore the modes well in low dimensions, they get completely stuck in one mode for $D \geq 16$. MAGSS and Meta-MAGSS with the Inverse Monge metric ($\alpha = 0.1$) are able to jump between the modes even for higher dimensions and the *meta-sampler* is overall the most accurate method.

Table 1: Mixture of narrow distributions.

sampler	metric	jump%	t(s)
PT	NA	6.18	2
DiGS	NA	0.27	2
MAGSS	$\boldsymbol{G}_{Ig}$, $\lambda = 1.0$	0.81	178
Meta-MAGSS	$\boldsymbol{G}_{IM}$, $\alpha^2 = 10^{-4}$	2.62	1327

Experiment specifications: For MALA we find a step-size of 60% acceptance rate for each dimension, since it is close to the optimal for Gaussians [Roberts and Rosenthal, 1998]. For MAGSS we try the Euclidean metric and the grid: $\alpha^2 \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ and $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$. We find $\alpha^2 = 0.1$ is always optimal. For Meta-MAGSS we fix $\alpha^2 = 0.1$ based on what was observed for MAGSS. We use Dopri5 solver with adaptive step-size. DiGS uses 10 MALA steps per iterations, $T = 100$ equally spaced times between $\alpha_1 = 10^{-4}$, $\alpha_{1000} = 1 - 10^{-4}$. PT uses $N = 100$ temperatures in the scale $\tau_i = b_{\min}^{-i/N}$ for $i = 1, .., N$ where $b_{\min} = 10^{-4}$.

5.3 MULTIMODAL WITH COMPLEX MODES

To demonstrate that we can simultaneously handle multimodality and complex local geometry, we consider a (uniform) mixture of two narrow bivariate distributions, the Rosenbrock and Squiggle distributions (Figure 6 left; the red line is purely for identifying jumps between the modes, Table 1). We use use the Inverse Monge and Inverse Generative metrics. However, Figure 6 (right) indicates that PT is the least accurate method, requiring substantially more samples for matching the target well. All methods will reach approximately the same Wasserstein distance if ran long enough, but both of our variants achieve it in less samples, confirming more efficient mixing.

Experiment specifications: Obtain $10,000$ samples using 10 chains. We run DiGS with 5 noise steps between 0.1 and 0.9, and 10 MALA iterations per sample. PT uses $\tau \in \{1.0, 5.62, 31.62, 177.83, 1000\}$ and thinning of 10. MAGSS and Meta-MAGSS are tuned using the same grid of values as Experiment 5.2, reporting the best based on distances. Meta-MAGSS uses 5 sweeps and 10 MALA iterations per sample. PT, DiGS and Meta-MAGSS rely on MALA with stepsize 0.001 ($\approx 60\%$ acceptance rate). We use $w = 3$ and $m = 8$ and the adaptive Dopri8 integrator.

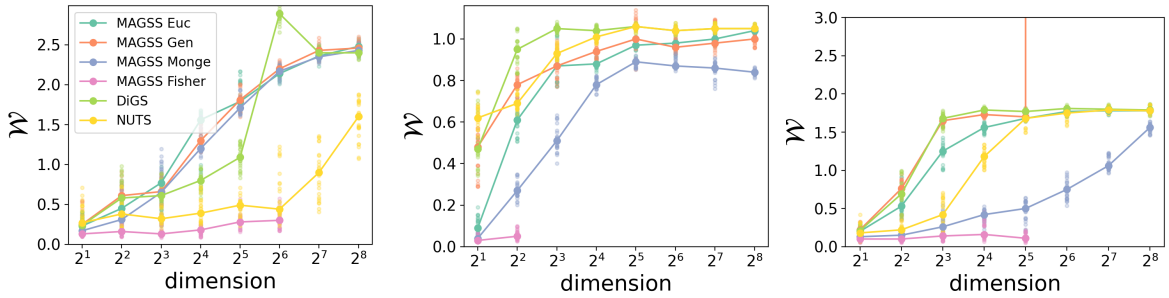

Figure 4: Univariate sampling accuracy in various metrics (Wasserstein distance, lower is better) for targets of varying dimensionality. The medians over 5 runs are connected with a line. Left: Funnel. Middle: Rosenbrock. Right: Squiggle.

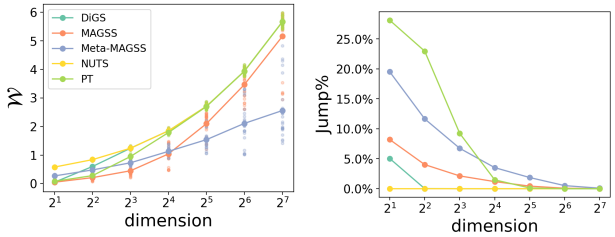

Figure 5: Accuracy (lower is better) for mixture of Gaussians. Both MAGSS variants use $\boldsymbol{G}_{IM}$ with $\alpha = 0.1$.

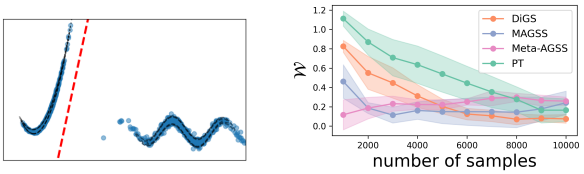

Figure 6: Left: Mixture of narrow distributions, with samples using Meta-MAGSS. Right: Wasserstein distance as a function of iterations (samples). For MAGSS we use $\boldsymbol{G}_{Ig}$ with $\lambda = 1$ and Meta-MAGSS $\boldsymbol{G}_{IM}$ with $\alpha^2 = 10^{-4}$.

5.4 FIELD SYSTEM

We include a highly multimodal target distribution of modes (2^D for $D = 16$) by replicating the Allen-Cahn Field System model experiment from Cabezas et al. [2024b]. The distribution has two global maxima at $(1, .., 1)$ and $(-1, .., -1)$, and several lower density modes at points of the form $x_i = \pm 1$ for all i. DiGS collapses to a single mode, PT explores only the two most dominant modes, while our Meta-MAGSS explores also the additional modes (Fig 7).

The target density is symmetric along each axis. The initial sampling position is $(-1, \ldots, -1)$ and we use the marginal distribution of x_8 to evaluate how well each sampler captures the symmetry. In particular, we report the percentage of samples with $x_8 > 0$, which should be 50% under the true distribution. Since reference samples are not directly available, we follow Cabezas et al. [2024b] and also report the Kernel Stein Discrepancy (KSD V-stat) [Liu et al., 2016] in Table 2.

Table 2: Field System model

sampler	KSD V-stat	$x_8 > 0$	t(s)
PT	0.13 ± 0.04	0.35	6
DiGS	0.13 ± 0.05	0.0	157
META-AGSS	2.98 ± 0.7	0.33	32

Our method explores more modes than the competing methods (PT and DiGS), although the KSD V-stat is worse. Note, however, that the KSD V-stat does not account for multimodality at all; DiGS has a better value despite covering only a single mode and failing completely in terms of the marginal distribution metric. In contrast, PT and Meta-MAGSS exhibit similar percentages of samples with $x_8 > 0$. We provide the density of the model, an explanation of the multimodality of the model, and the computation of KSD V-stat in Appendix C.5.

Experiment specifications: We obtain $10,000$ samples using 10 chains initialized at $(-1, .., -1)$. DiGS uses $T = 1000$ equally spaced times between $\alpha_1 = 10^{-5}$, $\alpha_{1000} = 1 - 10^{-5}$. PT uses $N = 200$ temperatures in the scale $\tau_i = b_{\min}^{-i/N}$ for $i = 1, .., N$ where $b_{\min} = 10^{-5}$. For Meta-MAGSS we try values of α and λ in powers of ten, finding $\lambda = 10^{-6}$ maximizes the number of jumps between modes and a single sweep. We use $w = 3$ and $m = 8$ and the Dopri5 integrator with adaptive step-size. For all methods MALA uses 10 iterations per sample and stepsize 0.005 (roughly 60% acceptance rate).

5.5 EFFECT OF NUMERICAL INTEGRATOR

We use numerical integrators for computing the geodesics in Eq. (2). To explore the effect of the integrator, we present results for broad range of integrators for a multimodal benchmark task considered previously by Chen et al. [2024]. The target is a 40-mode Gaussian mixture model with equal

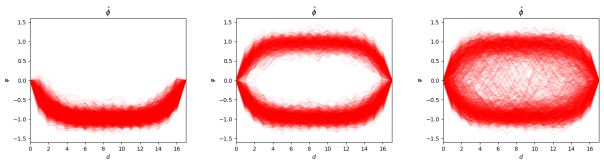

Figure 7: Samples from Allen-Cahn Field System model [Cabezas et al., 2024b] with 2^{16} modes that zig-zag between -1 and 1 on the y-axis over the $D = 16$ values at the x-axis. Euclidean DiGS (left) gets stuck in one mode, Parallel Tempering (middle) only explores two dominant modes with constant value over the x-axis, whereas Meta-MAGGS (right; $\boldsymbol{G}_{Ig}$ with $\lambda = 10^{-6}$) explores also the modes that switch between the extremes.

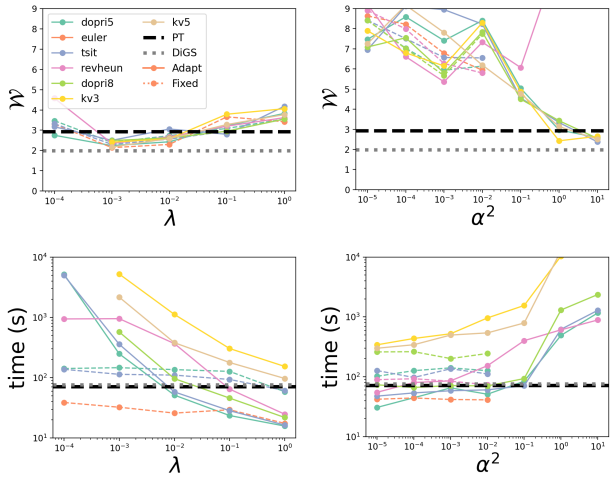

Figure 8: Different numerical integrators for the 40 mode Gaussian mixture model. The black dotted lines are PT and DiGS. Left: Inverse Generative metric with parameter λ. Right: Inverse Monge metric with parameter α^2.

weights and each component of variance $\sigma = 0.1$ where the means are distributed uniformly on the square $(-40, 40)^2$.

We use seven different integrators for $\boldsymbol{G}_{IM}$ and $\boldsymbol{G}_{Ig}$ metrics, including both adaptive and fixed step-sizes as implemented by Kidger [2021], and report the results in Figure 8. The three main conclusion are: (a) For metrics that are further away from Euclidean (large α or small λ) the integration time for adaptive methods grows dramatically. This is a natural consequence of operating in a less flat geometry. (b) For good accuracy we typically need to use such a geometry, which means there is inherent compromise between accuracy an computation. (c) Simple fixed-step integrators, even the Euler method, are efficient when they work, but for robustness we recommended adaptive methods. We recognize dopri5 as a good practical recommendation, but Euler is worth trying for the Inverse Generative metric.

Experiment specifications: Obtain $10,000$ samples using 10 chains. PT has $\tau \in \{1, 5.62, 31.62, 177.83, 1000\}$ and thinning of 200. DiGS uses $\alpha = 0.1$, thinning of 200 and 5 MALA steps per step. MALA has step size of 0.1. MAGSS is run with $w = 3, m = 8$ for the $\boldsymbol{G}_{IM}$ and $\boldsymbol{G}_{Ig}$ and metrics with the parameter grid of Experiment 5.2. We test seven different numerical integrators of Equation (2). The fixed integration size is 0.01. Details in Appendix B.

6 RELATED WORK

Lan et al. [2014] constructed the Wormhole Hamiltonian Monte Carlo where a specific geometry is built to (only) connect the modes of multimodal distributions. The modes are first identified along the Markov Chain evolution. After a new mode identification, it "stores" the mode's location for later use. A jump using the updated mode candidates guarantees correct detailed-balance equations. While this work served as an inspirational motivation for us, it requires notable additional components. However, MAGSS does not require separate identification or storage of the modes, but instead shrinks the distances by naturally warping the space.

7 CONCLUSIONS

Our aim was to show that local curvature and multimodality can be addressed by the same set of tools, namely Riemannian geometry. We provided a concrete Riemannian slice sampler, introduced two new metrics for improving mixing between modes, and showed that we can achieve accuracy and mixing comparable to recent samplers designed specifically for multimodal targets, by only using Riemannian metrics for this task.

One obvious limitation is the computational cost, caused by numeric integration of the geodesics. Even when using metrics with fast inverses, the per-iteration cost of MAGSS is larger than of competing methods. However, we note that we used maximally exact solvers rather than seeking for the highest computational efficiency. Now that the principle has been demonstrated, the use of more approximative numerical integrators for speeding up the overall computation could be studied in future work.

Acknowledgements

BW, HY, HPHL and AK were supported by the Research Council of Finland Flagship programme: Finnish Center for Artificial Intelligence (FCAI), and additionally by grants: 363317, 345811 and 348952. GA was supported by the DFF Sapere Aude Starting Grant "GADL". The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

Herbert Amann, Joachim Escher, and Gary Brookfield. *Analysis*. Springer, 2005.

Claude JP Bélisle, H Edwin Romeijn, and Robert L Smith. Hit-and-run algorithms for generating multivariate distributions. *Mathematics of Operations Research*, 18(2): 255–266, 1993.

Nils Berglund, Giacomo Di Gesu, and Hendrik Weber. An eyring-kramers law for the stochastic Allen-Cahn equation in dimension two. *Electronic Journal of Probability*, 22, 2017.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

Simon Byrne and Mark Girolami. Geodesic Monte Carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40(4):825–845, 2013.

Alberto Cabezas, Adrien Corenflos, Junpeng Lao, Rémi Louf, Antoine Carnec, Kaustubh Chaudhari, Reuben Cohn-Gordon, Jeremie Coullon, Wei Deng, Sam Duffield, et al. Blackjax: Composable Bayesian inference in Jax. *arXiv preprint arXiv:2402.10797*, 2024a.

Alberto Cabezas, Louis Sharrock, and Christopher Nemeth. Markovian flow matching: Accelerating MCMC with continuous normalizing flows. *arXiv preprint arXiv:2405.14392*, 2024b.

Wenlin Chen, Mingtian Zhang, Brooks Paige, José Miguel Hernández-Lobato, and David Barber. Diffusive Gibbs sampling. In *Forty-first International Conference on Machine Learning*, 2024.

Alain Durmus, Samuel Gruffaz, Mareike Hasenpflug, and Daniel Rudolf. Geodesic slice sampling on Riemannian manifolds. *arXiv preprint arXiv:2312.00417*, 2023.

David J Earl and Michael W Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

Charles J Geyer. Markov chain Monte Carlo maximum likelihood. *Computing science and statistics: proceedings of the 23rd symposium on the interface*, 1991.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Michael Habeck, Mareike Hasenpflug, Shantanu Kodgirwar, and Daniel Rudolf. Geodesic slice sampling on the sphere. *arXiv preprint arXiv:2301.08056*, 2023.

Marcelo Hartmann, Mark Girolami, and Arto Klami. Lagrangian manifold Monte Carlo on Monge patches. In *International Conference on Artificial Intelligence and Statistics*, pages 4764–4781. PMLR, 2022.

Marcelo Hartmann, Bernardo Williams, Hanlin Yu, Mark Girolami, Alessandro Barp, and Arto Klami. Warped geometric information on the optimisation of Euclidean functions, 2023. URL `https://arxiv.org/abs/2308.08305`.

Matthew D Hoffman, Andrew Gelman, et al. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

Patrick Kidger. *On neural differential equations*. PhD thesis, University of Oxford, 2021.

Beomsu Kim, Michael Puthawala, Jong Chul Ye, and Emanuele Sansone. (deep) generative geodesics. *arXiv preprint arXiv:2407.11244*, 2024.

Shiwei Lan, Jeffrey Streets, and Babak Shahbaba. Wormhole Hamiltonian Monte Carlo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

Shiwei Lan, Vasileios Stathopoulos, Babak Shahbaba, and Mark Girolami. Markov Chain Monte Carlo from Lagrangian Dynamics. *Journal of Computational and Graphical Statistics*, 24(2):357–378, 2015.

Krzysztof Łatuszyński and Daniel Rudolf. Convergence of hybrid slice sampling via spectral gap. *arXiv preprint arXiv:1409.2709*, 2014.

Krzysztof Łatuszyński, Matthew T Moores, and Timothée Stumpf-Fétizon. MCMC for multi-modal distributions. *arXiv preprint arXiv:2501.05908*, 2025.

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

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284. PMLR, 2016.

Iain Murray, Ryan Adams, and David MacKay. Elliptical slice sampling. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 541–548. JMLR Workshop and Conference Proceedings, 2010.

Viacheslav Natarovskii, Daniel Rudolf, and Björn Sprungk. Quantitative spectral gap estimate and Wasserstein contraction of simple slice sampling. *The Annals of Applied Probability*, 2021.

Radford M Neal. Slice sampling. *The annals of statistics*, 31(3):705–767, 2003.

Radford M Neal et al. MCMC using hamiltonian dynamics. *Handbook of markov chain Monte Carlo*, 2(11):2, 2011.

Filippo Pagani, Martin Wiegand, and Saralees Nadarajah. An n-dimensional Rosenbrock distribution for Markov chain Monte Carlo testing. *Scandinavian Journal of Statistics*, 49(2):657–680, 2022.

Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 1996.

Daniel Rudolf and Mario Ullrich. Comparison of hit-and-run, slice sampler and random walk Metropolis. *Journal of Applied Probability*, 55(4):1186–1202, 2018.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Robert H Swendsen and Jian-Sheng Wang. Replica Monte Carlo simulation of spin-glasses. *Physical review letters*, 57(21):2607, 1986.

Hakon Tjelmeland and Bjorn Kare Hegstad. Mode jumping proposals in MCMC. *Scandinavian journal of statistics*, 28(1):205–223, 2001.

Bernardo Williams, Hanlin Yu, Marcelo Hartmann, and Arto Klami. Geometric No-U-Turn samplers: Concepts and evaluation. In *12th International Conference on Probabilistic Graphical Models (PGM)*, pages 327–347. Journal of Machine Learning Research, 2024.

Dawn B Woodard, Scott C Schmidler, and Mark Huber. Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *The Annals of Applied Probability*, 2009.

Appendix

CONTENTS

A METRIC-AGNOSTIC GEODESIC SLICE SAMPLER

A.1 META-METRIC-AGNOSTIC GEODESIC SLICE SAMPLER

The Meta-MAGSSfound in algorithm 2 is the combination of MAGSS for K-steps followed by a local MCMC sampler for L-steps.

Algorithm 2 Meta-MAGSS

Input: Initial position $\boldsymbol{x}^{[0]}$ and metric components $\boldsymbol{G}(\boldsymbol{x})$. Parameters $m \in \mathbb{N}$, $w \geq 0$, K sweeps, L steps of local MCMC sampler.
Output: N samples $\boldsymbol{x}^{[n]}$.

for $n \leftarrow 1, \ldots, N$ **do**
  Let $\boldsymbol{x} \leftarrow \boldsymbol{x}^{[n-1]}$
  for $k \leftarrow 1, \ldots, K$ **do**
    Update $\boldsymbol{x}$ by MAGSS with initial position $\boldsymbol{x}$
  end for
  for $l \leftarrow 1, \ldots, L$ **do**
    Update $\boldsymbol{x}$ by local MCMC with initial position at $\boldsymbol{x}$.
  end for
  Set $\boldsymbol{x}^{[n]} \leftarrow \boldsymbol{x}$
end for

A.2 STEP-OUT AND SHRINKAGE PROCEDURES

The stepping-out and shrinkage procedures are Algorithm 3 and Algorithm 4 respectively, these algorithms are taken from Durmus et al. [2023]. Our code implementation of the step-out procedure has vectorized both while loops in Algorithm 3. This is done by evaluating the log density on all possible step-out points at once (vectorized). The code implementation of the shrinkage procedure (Algorithm 4) has a max number of iteration set at 100 for the while loop, which if exceeded defaults back to the previous point of the chain. In the algorithm boxes we use the notation for the exponential map $\gamma_{(x,v)}(t)$. JAX is used to handle automatic differentiation and the samplers are coded in the style of Blackjax [Bradbury et al., 2018, Cabezas et al., 2024a].

Algorithm 3 Stepping-out procedure. Call it Step-out$_{w,m}(s, \gamma_{(\boldsymbol{x},\boldsymbol{v})})$

Input: point $x \in \mathcal{M}$, direction $v \in \mathbb{S}_x^{d-1}$, level $s \in (0, p(x))$, hyperparameters $w \in (0, \infty)$ and $m \in \mathbb{N}$
Output: two points $\ell, r \in \mathbb{R}$ such that $\ell < 0 < r$

Draw $u \sim \text{Unif}([0, w])$.
Set $\ell := -u$ and $r := \ell + w$.
Draw $\iota \sim \text{Unif}(\{1, \dots, m\})$.
Set $i = 2$ and $j = 2$.
while $i \leq \iota$ and $p_{\mathcal{H}}(\gamma_{(x,v)}(\ell)) > s$ **do**
    Set $\ell = \ell - w$.
    Update $i = i + 1$.
end while
while $j \leq m + 1 - \iota$ and $p_{\mathcal{H}}(\gamma_{(x,v)}(r)) > s$ **do**
    Set $r = r + w$.
    Update $j = j + 1$.
end while
return (ℓ, r)

Algorithm 4 Shrinkage procedure. Call as Shrink$_{\ell,r}(s, \gamma_{(\boldsymbol{x},\boldsymbol{v})})$

Input: point $x \in \mathcal{M}$, direction $v \in \mathbb{S}_x^{d-1}$, level $s \in (0, p(x))$ and parameters $\ell < 0 < r$
Output: point $\theta \in L(x, v, s) \cap [\ell, r]$

Draw $\theta_h \sim \text{Unif}((0, r - l))$.
Set $\theta := \theta_h - 1_{\{\theta_h > r\}}(r - l)$.
Set $\theta_{\min} := \theta_h$.
Set $\theta_{\max} := \theta_h$.
while $p_{\mathcal{H}}(\gamma_{(x,v)}(\theta)) \leq s$ **do**
    if $\theta_h \in [\theta_{\min}, r - l]$ **then**
        Set $\theta_{\min} = \theta_h$.
    else
        Set $\theta_{\max} = \theta_h$.
    end if
    Draw $\theta_h \sim \text{Unif}((0, \theta_{\max}) \cup [\theta_{\min}, r - l))$.
    Set $\theta = \theta_h - 1_{\{\theta_h > r\}}(r - l)$.
end while
return θ.

A.3 THE GEODESIC EQUATIONS

A Riemannian metric is a smooth, symmetric, and positive-definite tensor $g : T_{\boldsymbol{x}}\,\mathcal{M} \times T_{\boldsymbol{x}}\,\mathcal{M} \to \mathbb{R}$ for each point $\boldsymbol{x} \in \mathcal{M}$. In coordinates, the metric is represented by a positive-definite matrix $\boldsymbol{G}(\boldsymbol{x})$ such that for all $\boldsymbol{v}, \boldsymbol{u} \in T_{\boldsymbol{x}}\,\mathcal{M}$,

$$g(\boldsymbol{v}, \boldsymbol{u}) = \boldsymbol{v}^\top \boldsymbol{G}(\boldsymbol{x})\boldsymbol{u}.$$

Geodesics are curves $\gamma(t)$ on $\mathcal{M}$ that locally minimize distance and generalize straight lines to curved spaces. They solve the geodesic equation, a second-order ODE determined by the metric. Given initial conditions $\gamma(0) = \boldsymbol{x}_0 \in \mathcal{M}$ and

$\dot{\gamma}(0) = \boldsymbol{v}_0 \in T_{\boldsymbol{x}_0}\mathcal{M}$, the geodesic equation in local coordinates is

$$\ddot{\gamma}^k(t) + \sum_{i,j=1}^{D} \Gamma_{ij}^k(\gamma(t))\dot{\gamma}^i(t)\dot{\gamma}^j(t) = 0, \quad \text{for } k = 1, \ldots, D,$$

where Γ_{ij}^k are the Christoffel symbols of the second kind, given by

$$\Gamma_{ij}^k = \tfrac{1}{2} \sum_{m=1}^{D} g^{km} \left(\partial_i g_{mj} + \partial_j g_{im} - \partial_m g_{ij}\right),$$

with $g_{ij} = \boldsymbol{G}(\boldsymbol{x})_{ij}$ and $g^{km} = (\boldsymbol{G}^{-1}(\boldsymbol{x}))_{km}$. Alternatively, defining the position-velocity system with $\boldsymbol{x} = \gamma(t)$ and $\boldsymbol{v} = \dot{\gamma}(t)$, the geodesic equations can be expressed as a first-order system (Equation 2):

$$\begin{aligned} \dot{\boldsymbol{x}}_k &= \boldsymbol{v}_k, \\ \dot{\boldsymbol{v}}_k &= -\|\boldsymbol{v}\|_{\Gamma^k}^2 \quad \text{for } k = 1, \ldots, D, \end{aligned}$$

where $\|\boldsymbol{v}\|_{\Gamma^k}^2 = \sum_{i,j=1}^{D} \Gamma_{ij}^k(\boldsymbol{x})\, \boldsymbol{v}_i\, \boldsymbol{v}_j$.

A.4 SAMPLING UNIFORMLY FROM THE UNIT TANGENT SPHERE

Recall the unit tangent sphere is defined by

$$\mathbb{S}_g^{D-1}(\boldsymbol{x}) := \{\boldsymbol{v} \in \mathbb{R}^D : \|\boldsymbol{v}\|_g^2 = 1\}.$$

Durmus et al. [2023, Appendix C.4] justify the existence of the uniform distribution distribution over $\mathbb{S}_g^{D-1}(\boldsymbol{x})$. One method for producing samples from the uniform distribution on the unit tangent sphere is:

1. Sample $\boldsymbol{z} \sim N(0, \boldsymbol{I})$.
2. Transform $\boldsymbol{v} \leftarrow \boldsymbol{G}^{-\frac{1}{2}}(\boldsymbol{x})\boldsymbol{z}$, then $\boldsymbol{v}$ is distributed according to $\mathcal{N}(0, \boldsymbol{G}^{-1}(\boldsymbol{x}))$.
3. Compute the Riemannian norm $\|\boldsymbol{v}\|_g = \sqrt{\boldsymbol{v}^T \boldsymbol{G}(\boldsymbol{x})\boldsymbol{v}}$.
4. Project to the boundary $\boldsymbol{v} \leftarrow \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_g}$.

A.5 THE HAUSDORFF MEASURE

The volume form of a Riemannian manifold with metric $\boldsymbol{G}(\boldsymbol{x})$ is defined as $V(\mathrm{d}\boldsymbol{x}) := \sqrt{\det \boldsymbol{G}(\boldsymbol{x})}\, \mathrm{d}\boldsymbol{x}$. For technical details about the volume form, interested readers can consult Proposition 2.41 in Lee [2018]. The volume element gives the natural measure on the manifold, analogous to the Lebesgue measure in Euclidean space [Durmus et al., 2023]. The Hausdorff density is defined as the density which integrates to one with respect to the volume element:

$$p_{\mathcal{H}}(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{\sqrt{\det \boldsymbol{G}(\boldsymbol{x})}}.$$

An intuitive explanation for the volume element can be thought in terms of change-of-variables in Euclidean space. When transforming coordinates via a diffeomorphism $\phi : \mathbb{R}^D \to \mathbb{R}^D$, the standard density must be adjusted by the Jacobian determinant to preserve probability mass. That is, $|\det J|$ accounts for local volume distortion.

When $\phi : \mathbb{R}^D \to \mathbb{R}^d$, where $d > D$ maps from a lower-dimensional Euclidean space onto a manifold embedded in higher dimensions, the Jacobian J is generally rectangular. In this case, the induced Riemannian metric (Pullback metric) on the manifold is $\boldsymbol{G}(\boldsymbol{x}) = JJ^\top$, and the volume change is given by $\sqrt{\det \boldsymbol{G}(\boldsymbol{x})}$, which generalizes $|\det J|$.

Thus, the Hausdorff density $p_{\mathcal{H}}$ adjusts the density p to be properly normalized on the manifold with respect to the intrinsic geometry. This adjustment ensures correct sampling and integration as seen in Proposition 1.

A.6 OBSERVATIONS AND PROPOSITION

Denote by $\boldsymbol{G}_{IM}(\boldsymbol{x})$ the inverse Monge metric and by $\boldsymbol{G}_{Ig}(\boldsymbol{x})$ the inverse generative metric. The metrics are defined as:

$$\boldsymbol{G}_{Ig}(\boldsymbol{x}) = \left(\frac{p(\boldsymbol{x})+\lambda}{p_0+\lambda}\right)^2 \boldsymbol{I}, \quad \boldsymbol{G}_{IM}(\boldsymbol{x}) = \boldsymbol{I} - \frac{\alpha^2}{1+\alpha^2\|\nabla\ell(\boldsymbol{x})\|^2}\nabla\ell(\boldsymbol{x})\nabla\ell(\boldsymbol{x})^\top.$$

Observation 1. Let $p(\boldsymbol{x})$ be a smooth density function. Let $(\boldsymbol{x}_t, \boldsymbol{v}_t)$ be the geodesic flow with initial conditions $(\boldsymbol{x}_0, \boldsymbol{v}_0)$ such that $p(x_0) > 0$ with respect to the Inverse Generative metric. Then, for t such that $p(\boldsymbol{x}_t) \to 0$ we have $\|\boldsymbol{v}_t\|_2 > \|\boldsymbol{v}_t\|_0$.

Analysis for the Inverse Generative metric Assume a geodesic curve starting at $(\boldsymbol{x}_0, \boldsymbol{v}_0)$ satisfies $p(\boldsymbol{x}_0) \geq p(\boldsymbol{x}_t)$ for all $t \geq 0$ and $p(\boldsymbol{x}_t) \to 0$. Recall that along a geodesic curve, the magnitude of the velocity with respect to the metric remains constant:

$$\|\boldsymbol{v}_t\|^2_{\boldsymbol{G}_{Ig}} = \|\boldsymbol{v}_0\|^2_{\boldsymbol{G}_{Ig}} \quad \forall t.$$

Thus, the equality holds:

$$\begin{aligned}
\|\boldsymbol{v}_t\|^2_{\boldsymbol{G}_{Ig}} &= \|\boldsymbol{v}_0\|^2_{\boldsymbol{G}_{Ig}} \\
\left(\frac{p(\boldsymbol{x}_0)+\lambda}{p_0+\lambda}\right)^2 \|\boldsymbol{v}_0\|^2 &= \left(\frac{p(\boldsymbol{x}_t)+\lambda}{p_0+\lambda}\right)^2 \|\boldsymbol{v}_t\|_2^2 \\
\left(\frac{p(\boldsymbol{x}_0)+\lambda}{p(\boldsymbol{x}_t)+\lambda}\right)^2 \|\boldsymbol{v}_0\|^2 &= \|\boldsymbol{v}_t\|_2^2.
\end{aligned}$$

Since $p(\boldsymbol{x}_0) \geq p(\boldsymbol{x}_t)$, it follows that $\|\boldsymbol{v}_t\|_2^2 \geq \|\boldsymbol{v}_0\|_2^2$, and as $p(\boldsymbol{x}_t) \to 0$ the quantity is arbitrary large.

Observation 2. Let $p(\boldsymbol{x})$ be a smooth density function. Let $(\boldsymbol{x}_t, \boldsymbol{v}_t)$ be the geodesic flow with initial conditions $(\boldsymbol{x}_0, \boldsymbol{v}_0)$ with respect to the Inverse Monge metric, such that $\boldsymbol{x}_0$ is a local maximum. Then $\|\boldsymbol{v}_t\|_2 \geq \|\boldsymbol{v}_0\|_2$ for all $t \neq 0$.

Analysis for the Inverse Monge metric We consider geodesics emanating from a mode. Let $\boldsymbol{x}_0$ be a mode, meaning that $\nabla\ell(\boldsymbol{x}_0) = 0$. Consider a geodesic emanating form $\boldsymbol{x}_0$ with velocity $\boldsymbol{v}_0$, it holds

$$\|\boldsymbol{v}_0\|_2^2 - \frac{\alpha^2}{1+\alpha^2\|\nabla\ell(\boldsymbol{x}_0)\|^2}\langle\nabla\ell(\boldsymbol{x}_0), \boldsymbol{v}_0\rangle^2 = \|\boldsymbol{v}_t\|_2^2 - \frac{\alpha^2}{1+\alpha^2\|\nabla\ell(\boldsymbol{x}_t)\|^2{}_2}\langle\nabla\ell(\boldsymbol{x}_t), \boldsymbol{v}_t\rangle^2,$$

or

$$\|\boldsymbol{v}_0\|_2^2 + \frac{\alpha^2}{1+\alpha^2\|\nabla\ell(\boldsymbol{x}_t)\|^2{}_2}\langle\nabla\ell(\boldsymbol{x}_t), \boldsymbol{v}_t\rangle^2 = \|\boldsymbol{v}_t\|_2^2.$$

We see that $\| \boldsymbol{v}_t \|^2 \geq \| \boldsymbol{v}_0 \|^2$, so any geodesic starting from $\boldsymbol{x}_0$ always has a "shrinkage" behavior. So this metric helps bring the entire space close to $\boldsymbol{x}_0$ along geodesics, as it shrinks the space towards (multiple) modes. Also note that the collapsing towards modes depends on how flat the region is and how well-aligned the velocity and the gradient of ℓ is. For complicated distributions, the behavior should not depend monotonically on α.

Note: For a multimodal distribution of dim ≥ 2, Observation 1 and Observation 2 guarantee that low-density/increasing-gradient regions the speed increases, but we do not have the guarantee that the geodesics given by the inverse metrics will reach the other modes. The geodesic could twist before reaching the other modes, which could negate the "teleport/move fast" effect.

Proposition 1. An MCMC sampler targeting the Hausdorff density on a Riemannian manifold $\mathcal{M}$ with metric tensor $\boldsymbol{G}(\boldsymbol{x})$ also targets the correct distribution on the Euclidean space.

For a general setting proof of the Proposition consult Section XII.1, Proposition 1.5 in Amann et al. [2005].

The volume element on the manifold is defined as $V(\mathrm{d}\boldsymbol{x}) = \sqrt{\det \boldsymbol{G}(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}$, where $\boldsymbol{G}(\boldsymbol{x})$ is the Riemannian metric tensor. Let $\boldsymbol{X}$ be a random variable on $\mathcal{M}$ whose law is $p_{\mathcal{H}}$ where $p_{\mathcal{H}}(\boldsymbol{x})$ is the Hausdorff density. Let $B \in \mathcal{B}(\mathcal{M})$ be a Borel set on the manifold $\mathcal{M}$. The probability of $\boldsymbol{X}$ being in B, under the Hausdorff target density, is given by

$$\mathbb{P}(\boldsymbol{X} \in B) = \int_B p_{\mathcal{H}}(\boldsymbol{x}) V(\mathrm{d}\boldsymbol{x}). \tag{1}$$

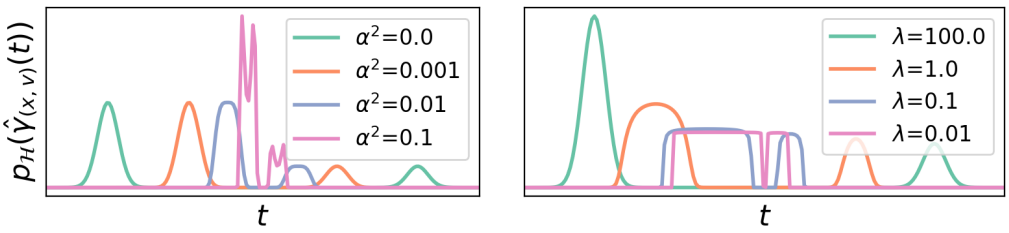

Figure 1: The Hausdorff density of a mixture of two Gaussian distributions evaluated along the geodesic, namely $t \mapsto p_{\mathcal{H}}(\hat{\gamma}_{(\boldsymbol{x},\boldsymbol{v})}(t))$ for different values of λ and α^2. Left: inverse Monge metric. Right: inverse Generative metric

Substituting $p_{\mathcal{H}}(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{\sqrt{\det \boldsymbol{G}(\boldsymbol{x})}}$,

$$\begin{aligned}\mathbb{P}(\boldsymbol{X} \in B) &= \int_B \frac{p(\boldsymbol{x})}{\sqrt{\det G(\boldsymbol{x})}} \sqrt{\det \boldsymbol{G}(\boldsymbol{x})}\, \mathrm{d}\boldsymbol{x} \\ &= \int_B p(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}\,. \end{aligned} \tag{2}$$

Thus, the integral of the Hausdorff density with respect to the volume element on the manifold coincides with the integral of the Euclidean density $p(\boldsymbol{x})$ over the same set B.

Since the probabilities computed for any $B \in \mathcal{B}(\mathcal{M})$ are identical whether using (1) or (2), the corresponding estimators for the probabilities also coincide. Consequently, an MCMC sampler on the manifold targeting the Hausdorff density $p_{\mathcal{H}}(\boldsymbol{x})$ correctly targets the Euclidean density $p(\boldsymbol{x})$ in $\mathbb{R}^D$.

B ADDITIONAL EXPERIMENTAL RESULTS

B.1 LOGISTIC REGRESSION

Here we denote by θ the random variable of interest and by the x input data. The Logistic regression model [Girolami and Calderhead, 2011] is

$$p(\boldsymbol{y}_i \,|\, \boldsymbol{\theta}, \boldsymbol{x}_i) = \text{Bernoulli}(\boldsymbol{y}_i \,|s(\boldsymbol{x}_i^\top\, \boldsymbol{\theta})), \quad p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}\,|0, \alpha\, \boldsymbol{I}_D), \quad i = 1, .., N,$$

where $\alpha = 100$ and $s(\cdot)$ is the Sigmoid function. The Fisher Information Metric for this probabilistic model including the addition of the the Hessian of the prior is: $\boldsymbol{G}(\boldsymbol{x}) = \boldsymbol{X}^\top\, \boldsymbol{\Lambda}\, \boldsymbol{X} + \alpha^{-1}\, \boldsymbol{I}$. Where $\boldsymbol{X}$ is the covariate matrix and $\boldsymbol{\Lambda}$ is a diagonal matrix with entries $\boldsymbol{\Lambda}_{nn} = s(\boldsymbol{x}_i^\top\, \boldsymbol{\theta})\big(1 - s(\boldsymbol{x}_i^\top\, \boldsymbol{\theta})\big)$.

References samples used for computing the Wasserstein distance are obtained with HMC-NUTS. The samples obtained with the Euclidean and Fisher metrics are just as close to the samples, but Fisher and Monge have higher effective sample size (ESS) and use less shrinkage iterations than the Euclidean metric (See Table 1). The Monge metric uses the parameter $\alpha = 1$. $\mathcal{W}$ is the earth mover's distance. The notation used is $\text{mean} \pm \text{std}$ over 5 runs with different seeds.

B.2 NUMERICAL INTEGRATORS

The numerical solvers we consider are part of the `diffrax` package [Kidger, 2021]. We consider three groups of solvers. Simple solvers (euler, tsit, dopri). Implicit solvers (kv). And reversible solvers (revheun). The solvers have the following characteristics:

- euler: The Euler solver can only be used with a fixed step-size.
- tsit: Tsitouras' 5/4 method can be used with both fixed and adaptive step-size.
- dopri5: Dormand-Prince's 5/4 method can be used with both fixed and adaptive step-size.
- dopri8: Dormand-Prince's 8/7 method can be used with both fixed and adaptive step-size.

model	metr	$\mathcal{W}$	min ESS	avg ESS	avg step-out	avg shrinkage	t(s)
aus	euclidean	0.74 ± 0.12	18 ± 5	228 ± 12	0.14 ± 0.0	3.33 ± 0.01	8.2
	fisher	0.58 ± 0.01	177 ± 9	270 ± 13	0.95 ± 0.0	1.0 ± 0.01	3298.6
ger	euclidean	0.49 ± 0.01	35 ± 10	119 ± 9	0.08 ± 0.0	4.19 ± 0.02	15.8
	monge	0.75 ± 0.02	80 ± 5	4673 ± 113	3.97 ± 0.03	0.28 ± 0.01	4665.4
	fisher	0.49 ± 0.0	85 ± 24	169 ± 5	0.95 ± 0.0	0.99 ± 0.01	19684.4
hrt	euclidean	0.63 ± 0.01	137 ± 26	254 ± 16	0.19 ± 0.01	2.72 ± 0.02	7.3
	monge	0.74 ± 0.03	394 ± 55	2979 ± 266	2.99 ± 0.04	0.38 ± 0.02	637.4
	fisher	0.64 ± 0.01	232 ± 14	311 ± 8	0.92 ± 0.01	1.01 ± 0.01	1694.8
pim	euclidean	0.21 ± 0.0	266 ± 41	445 ± 45	0.11 ± 0.0	3.59 ± 0.03	6.8
	monge	0.29 ± 0.01	515 ± 72	4656 ± 269	2.38 ± 0.03	0.53 ± 0.02	1711.7
	fisher	0.21 ± 0.0	427 ± 47	547 ± 30	0.93 ± 0.0	0.98 ± 0.01	425.4
rip	euclidean	0.09 ± 0.01	829 ± 112	1499 ± 56	0.24 ± 0.01	2.46 ± 0.02	4.2
	monge	0.15 ± 0.02	1042 ± 211	3123 ± 685	1.38 ± 0.02	0.97 ± 0.01	593.9
	fisher	0.09 ± 0.0	1623 ± 113	1753 ± 92	0.91 ± 0.0	0.96 ± 0.01	204.7

Table 1: Bayesian Logistic Regression. Entries are reported as $\mathrm{mean} \pm \mathrm{std}$.

- kv3: Kvaerno's 3/2 method is an implicit solver can be only used with adaptive step-size.
- kv5: Kvaerno's 5/4 method is an implicit solver can only be used adaptive step-size.
- revheun: Reversible Heun method can be used with both fixed and adaptive step-size.

C MATHEMATICAL DERIVATIONS

C.1 THE GENERATIVE AND INVERSE GENERATIVE METRICS

The Generative and Inverse Generative metrics read

$$G(x) = f(x)I = \exp(\log f(x))I,$$

where the scalar factor is $f(x) = \left(\frac{p_0+\lambda}{p(x)+\lambda}\right)^2$ for the Generative metric and $f(x) = \left(\frac{p(x)+\lambda}{p_0+\lambda}\right)^2$ for the Inverse Generative metric.

Square root and inverse square root The quantities are given by:

$$\begin{aligned} G^{\frac{1}{2}}(x) &= \exp\{\tfrac{1}{2}\log f(x)\}I, \\ G^{-\frac{1}{2}}(x) &= \exp\{-\tfrac{1}{2}\log f(x)\}I, \\ \log|\det G(\boldsymbol{x})| &= D\log f(x). \end{aligned}$$

Christoffel symbols derivation Given the Riemannian metric $G(x) = f(x)I$, the tensor entries are:

$$G_{ij}(x) = f(x)\delta_{ij}.$$

The Christoffel symbols for this metric are given by:

$$\begin{aligned} \Gamma_{ij}^k &= \frac{1}{2f(x)}\left(\delta_{jk}\partial_i f(x) + \delta_{ik}\partial_j f(x) - \delta_{ij}\partial_k f(x)\right) \\ &= \frac{1}{2}\left(\delta_{jk}\partial_i \log f(x) + \delta_{ik}\partial_j \log f(x) - \delta_{ij}\partial_k \log f(x)\right). \end{aligned}$$

Denote by e_k the standard basis vectors, the Christoffel symbols in matrix notation are:

$$\Gamma^k = \frac{1}{2}\left(\nabla \log f(x)\, e_k^\top + e_k\, \nabla \log f(x)^\top - \partial_k \log f(x) I\right).$$

We compute $\|\boldsymbol{v}\|^2_{\Gamma^k} = \sum_{i,j=1}^D \Gamma^k_{ij}(\boldsymbol{x})\, \boldsymbol{v}_i\, \boldsymbol{v}_j$ which appears in the geodesic equations,

$$\|\boldsymbol{v}\|^2_{\Gamma^k} = \langle \boldsymbol{v}, \nabla \log f \rangle\, \boldsymbol{v}_k - \tfrac{1}{2}\|\boldsymbol{v}\|^2 \partial_k \log f.$$

The geodesic equations read:

$$\begin{aligned}\dot{\boldsymbol{x}} &= \boldsymbol{v},\\ \dot{\boldsymbol{v}} &= \tfrac{1}{2}\|\boldsymbol{v}\|^2 \nabla \log f - \langle \boldsymbol{v}, \nabla \log f \rangle\, \boldsymbol{v}\,.\end{aligned}$$

C.2 THE MONGE AND INVERSE MONGE METRICS

The Monge metric and the Inverse Monge metric are:

$$G(x) = I + \alpha^2 \nabla\ell \nabla\ell^\top, \quad G^{-1}(x) = I - \frac{\alpha^2}{1+\alpha^2\|\nabla\ell\|^2}\nabla\ell\nabla\ell^\top.$$

Square root and inverse square root Define the quantity $L_\alpha := 1 + \alpha^2\|\nabla\ell\|^2$, we list the quantities derived from the matrix and present later their derivation,

$$\begin{aligned} G^{1/2}(x) &= I + \frac{\alpha^2}{1+\sqrt{L_\alpha}}\nabla\ell(\boldsymbol{x})\nabla\ell(\boldsymbol{x})^\top,\\ G^{-1/2}(x) &= I - \frac{\alpha^2}{L_\alpha+\sqrt{L_\alpha}}\nabla\ell(x)\nabla\ell(x)^\top,\\ \log|\det G(\boldsymbol{x})| &= \log\Big(1+\alpha^2\|\nabla\ell\|^2\Big).\end{aligned}$$

We now derive the quantities. Hartmann et al. [2022] gave

$$G^{-\frac{1}{2}}(x) = I + \frac{1}{\|\nabla\ell\|^2}\left(\frac{1}{\sqrt{L_\alpha}} - 1\right)\nabla\ell\nabla\ell^\top \tag{3}$$

Note that if $\|\nabla\ell\|^2 \to 0$ then Equation 3 is undefined. We find a more numerical stable form of $G^{-\frac{1}{2}}(x)$ Multiply the scalar $\frac{1}{\|\nabla\ell\|^2}\left(\frac{1}{\sqrt{L_\alpha}} - 1\right)$ by its conjugate

$$\frac{1}{\|\nabla\ell\|^2}\left(\frac{1}{\sqrt{L_\alpha}} - 1\right) = \frac{1}{\|\nabla\ell\|^2}\left(\frac{1-\sqrt{L_\alpha}}{\sqrt{L_\alpha}}\right)\left(\frac{1+\sqrt{L_\alpha}}{1+\sqrt{L_\alpha}}\right) = \frac{-\alpha^2}{L_\alpha+\sqrt{L_\alpha}}.$$

Plugging the scalar $\frac{-\alpha^2}{L_\alpha+\sqrt{L_\alpha}}$ into $G^{-\frac{1}{2}}(\boldsymbol{x})$, then it is numerically stable for $\|\nabla\ell\|^2 \to 0$.

The computation of $G^{\frac{1}{2}}(x)$ For convenience take $y = \nabla\ell(x)$. The the metric is $G(y) = I + yy^\top$. Let us assume the square root is of the form $G^{\frac{1}{2}}(y) = I + \lambda yy^\top$. Let us formulate a quadratic equation for λ:

$$\begin{aligned} G^{\frac{1}{2}}(y)G^{\frac{1}{2}}(y) &= I + yy^\top\\ I + 2\lambda yy^\top + \lambda^2\|y\|^2 yy^\top &= I + yy^\top\\ 0 &= \left(1 - 2\lambda - \lambda^2\|y\|^2\right)yy^\top.\end{aligned}$$

The solutions of the quadratic equation are

$$\lambda = \frac{-1 \pm \sqrt{1+\|y\|^2}}{\|y\|^2}.$$

Let us simplify $\frac{-1+\sqrt{1+\|y\|^2}}{\|y\|^2}$, multiply by its conjugate

$$\frac{\sqrt{1+\|y\|^2}-1}{\|y\|^2}\left(\frac{\sqrt{1+\|y\|^2}+1}{\sqrt{1+\|y\|^2}+1}\right) = \frac{\|y\|^2}{\|y\|^2\sqrt{1+\|y\|^2}+1}.$$

Substitute $y = \alpha\nabla\ell(\boldsymbol{x})$ and we obtain the result

$$G^{1/2}(x) = I + \frac{\alpha^2}{\sqrt{1+\alpha^2\|\nabla\ell(\boldsymbol{x})\|^2}+1}\nabla\ell\nabla\ell^\top.$$

Christoffel symbols of the Monge metric The Christoffel associated to the Monge metric (derivation in Section C.3 and Hartmann et al. [2022]) are

$$\Gamma^k(x) = \frac{\alpha^2}{1+\alpha^2\|\nabla\ell\|^2}\nabla^2\ell\partial_k\ell,$$

and the geodesic equations read

$$\begin{aligned}\dot{\boldsymbol{x}} &= \boldsymbol{v},\\ \dot{\boldsymbol{v}} &= -\frac{\alpha^2}{L_\alpha}\|\boldsymbol{v}\|^2_{\nabla^2\ell}\nabla\ell.\end{aligned}$$

Christoffel symbols of the Inverse Monge metric The Christoffel symbols associated to the inverse Monge metric (derivation in Section C.4) are

$$\Gamma^k = \frac{\alpha^2}{2}\Big[L_\alpha\left(\nabla f\nabla\ell^\top + \nabla\ell\nabla f^\top + 2f\nabla^2\ell\right)\partial_k\ell + \nabla\ell\nabla\ell^\top\partial_k f + \alpha^2\left\langle\nabla\ell,\nabla f\right\rangle\nabla\ell\nabla\ell^\top\partial_k\ell\Big],$$

and the geodesic equations read

$$\begin{aligned}\dot{\boldsymbol{x}} &= \boldsymbol{v},\\ \dot{\boldsymbol{v}} &= -\frac{\alpha^2}{2}\Big[2L_\alpha\left(\langle v,\nabla f\rangle\langle\nabla\ell,v\rangle + f\|v\|^2_{\nabla^2\ell}\right)\nabla\ell + \langle\nabla\ell,v\rangle^2\nabla f + \alpha^2\left\langle\nabla\ell,\nabla f\right\rangle\langle\nabla\ell,v\rangle^2\nabla\ell\Big].\end{aligned}$$

C.3 MONGE METRIC: CHRISTOFFEL SYMBOLS DERIVATION

For completeness let us do an alternative derivation of the Christoffel symbols from the one found in Hartmann et al. [2022]. Take the auxiliary function $f(x) = -\frac{1}{L_\alpha}$, where $L_\alpha = 1+\alpha^2\|\nabla\ell\|^2$. The metric and inverse components are:

$$\begin{aligned}g_{ij} &= \delta_{ij} + \alpha^2\partial_i\ell\partial_j\ell,\\ g^{km} &= \delta_{km} + \alpha^2 f(x)\partial_k\ell\partial_m\ell.\end{aligned}$$

The derivatives of the metric are:

$$\begin{aligned}\partial_i g_{mj} &= \alpha^2\left(\partial_{im}\ell\partial_j\ell + \partial_m\ell\partial_{ij}\ell\right),\\ \partial_j g_{im} &= \alpha^2\left(\partial_{ij}\ell\partial_m\ell + \partial_i\ell\partial_{jm}\ell\right),\\ \partial_m g_{ij} &= \alpha^2\left(\partial_{im}\ell\partial_j\ell + \partial_i\ell\partial_{jm}\ell\right).\end{aligned}$$

The Christoffel symbols read,

$$\begin{aligned}\Gamma^k_{ij} &= \frac{1}{2}g^{km}\left(\partial_i g_{mj} + \partial_j g_{im} - \partial_m g_{ij}\right)\\ &= \frac{\alpha^2}{2}g^{km}\left(2\partial_m\ell\partial_{ij}\ell\right)\\ &= \alpha^2\sum_m\left(\delta_{km} + \alpha^2 f(x)\partial_k\ell\partial_m\ell\right)\partial_m\ell\partial_{ij}\ell\\ &= \alpha^2\left(\partial_k\ell\partial_{ij}\ell + \alpha^2 f(x)\sum_m(\partial_m\ell)^2\partial_k\partial_{ij}\ell\right)\\ &= \alpha^2\partial_k\ell\partial_{ij}\left(1 - \tfrac{\alpha^2\|\nabla\ell\|^2}{1+\alpha^2\|\nabla\ell\|^2}\right)\\ &= \frac{\alpha^2}{1+\alpha^2\|\nabla\ell\|^2}\partial_k\ell\partial_{ij}.\end{aligned}$$

Thus, the Christoffel symbols are $\Gamma^k_{ij} = \frac{\alpha^2}{L_\alpha}\partial_k\ell\partial_{ij}$. Writing in matrix form Γ^k of size $D\times D$ with components $[\Gamma^k]_{ij} = \Gamma^k_{ij}$,

$$\Gamma^k = \frac{\alpha^2}{1+\alpha^2\|\nabla\ell\|^2}\nabla^2\ell\partial_k\ell.$$

Let us compute $\|\boldsymbol{v}\|^2_{\Gamma^k}$, which appears in the geodesic equations,

$$v^\top \Gamma^k v = \frac{\alpha^2}{1+\alpha^2\|\nabla\ell\|^2}\|v\|^2_{\nabla^2\ell}\partial_k\ell.$$

The geodesic equations read:

$$\begin{aligned}\dot{\boldsymbol{x}} &= \boldsymbol{v},\\ \dot{\boldsymbol{v}} &= -\frac{\alpha^2}{L_\alpha}\|\boldsymbol{v}\|^2_{\nabla^2\ell}\nabla\ell.\end{aligned}$$

C.4 INVERSE MONGE METRIC: CHRISTOFFEL SYMBOLS DERIVATION

Again the auxilary function is $f(x) = -\frac{1}{L_\alpha}$, where $L_\alpha = 1+\alpha^2\|\nabla\ell\|^2$, the metric and inverse components are

$$\begin{aligned}g_{ij} &= \delta_{ij} + f(x)\alpha^2\partial_i\ell\partial_j\ell\\ g^{km} &= \delta_{km} + \alpha^2\partial_k\ell\partial_m\ell.\end{aligned}$$

The derivatives of the metric are (we mark with the same color repeating terms)

$$\begin{aligned}\partial_i g_{mj} &= \alpha^2\left(\partial_i f\partial_m\ell\partial_j\ell + f\partial_{im}\ell\partial_j\ell + f\partial_m\ell\partial_{ij}\ell\right),\\ \partial_j g_{im} &= \alpha^2\left(\partial_j f\partial_i\ell\partial_m\ell + f\partial_{ij}\ell\partial_m\ell + f\partial_i\ell\partial_{jm}\ell\right),\\ \partial_m g_{ij} &= \alpha^2\left(\partial_m f\partial_i\ell\partial_j\ell + f\partial_{mi}\ell\partial_j\ell + f\partial_i\ell\partial_{mj}\ell\right).\end{aligned}$$

Let us compute the Christoffel symbols of the first kind (blue and red terms will cancel out, pink terms add to each other)

$$\begin{aligned}\Gamma_{kij} &= \frac{1}{2}\left(\partial_i g_{mj} + \partial_j g_{im} - \partial_m g_{ij}\right)\\ &= \frac{\alpha^2}{2}\left(\partial_i f\partial_m\ell\partial_j\ell + \partial_j f\partial_i\ell\partial_m\ell - \partial_m f\partial_i\ell\partial_j\ell + 2f\partial_{ij}\ell\partial_m\ell\right).\end{aligned}$$

The Christoffel symbols of the second kind read

$$\begin{aligned}\Gamma^k_{ij} &= \frac{1}{2}g^{km}\left(\partial_i g_{mj} + \partial_j g_{im} - \partial_m g_{ij}\right)\\ &= \frac{\alpha^2}{2}\sum_m\left(\delta_{km} + \alpha^2\partial_k\ell\partial_m\ell\right)\left(\partial_i f\partial_m\ell\partial_j\ell + \partial_j f\partial_i\ell\partial_m\ell - \partial_m f\partial_i\ell\partial_j\ell + 2f\partial_{ij}\ell\partial_m\ell\right)\\ &= \frac{\alpha^2}{2}\Bigg(\partial_i f\partial_k\ell\partial_j\ell + \partial_j f\partial_i\ell\partial_k\ell - \partial_k f\partial_i\ell\partial_j\ell + 2f\partial_{ij}\ell\partial_k\ell\\ &\quad + \alpha^2\partial_k\ell\left(\partial_i f\|\nabla\ell\|^2\partial_j\ell + \partial_j f\partial_i\ell\|\nabla\ell\|^2 - \langle\nabla f,\nabla\ell\rangle\,\partial_i\ell\partial_j\ell + 2f\partial_{ij}\ell\|\nabla\ell\|^2\right)\Bigg)\\ &= \frac{\alpha^2}{2}\Bigg(\partial_k\ell\left(L_\alpha\partial_i f\partial_j\ell + L_\alpha\partial_i\ell\partial_j f - \alpha^2\langle\nabla f,\nabla\ell\rangle\,\partial_i\ell\partial_j\ell + 2L_\alpha f(x)\partial_{ij}\ell\right) - \partial_k f\partial_i\ell\partial_j\ell\Bigg).\end{aligned}$$

Thus, the Christoffel symbols are:

$$\Gamma^k_{ij} = \frac{\alpha^2}{2}\Big[\partial_k\ell\left(L_\alpha\left(\partial_i f\partial_j\ell + \partial_i\ell\partial_j f + 2f(x)\partial_{ij}\ell\right) - \alpha^2\langle\nabla f,\nabla\ell\rangle\,\partial_i\ell\partial_j\ell\right) - \partial_k f\partial_i\ell\partial_j\ell\Big].$$

Written in matrix form

$$\Gamma^k = \frac{\alpha^2}{2}\Big[\partial_k\ell\left(L_\alpha\left(\nabla f\nabla\ell^\top + \nabla\ell\nabla f^\top + 2f(x)\nabla^2\ell\right) - \alpha^2\langle\nabla f,\nabla\ell\rangle\,\nabla\ell\nabla\ell^\top\right) - \partial_k f\nabla\ell\nabla\ell^\top\Big].$$

Let us compute $\|v\|^2_{\Gamma^k}$ which appears in the geodesic equations

$$v^\top \Gamma^k v = \frac{\alpha^2}{2}\Big[\partial_k \ell \left(2L_\alpha \left(\langle v, \nabla f\rangle \langle \nabla \ell, v\rangle + f\|v\|^2_{\nabla^2 \ell}\right) - \alpha^2 \langle \nabla f, \nabla \ell\rangle \langle \nabla \ell, v\rangle^2\right) - \partial_k f \langle \nabla \ell, v\rangle^2\Big].$$

Then the geodesic equations read,

$$\begin{aligned}
\dot{\boldsymbol{x}} &= \boldsymbol{v}, \\
\dot{\boldsymbol{v}} &= -\frac{\alpha^2}{2}\Big[\left(2L_\alpha \left(\langle v, \nabla f\rangle \langle \nabla \ell, v\rangle + f\|v\|^2_{\nabla^2 \ell}\right) - \alpha^2 \langle \nabla f, \nabla \ell\rangle \langle \nabla \ell, v\rangle^2\right) \nabla \ell - \langle \nabla \ell, v\rangle^2 \nabla f\Big].
\end{aligned}$$

Where the gradient of f is:

$$\nabla f = \frac{2\alpha^2}{L_\alpha^2} \nabla^2 \ell \nabla \ell.$$

C.5 TARGET DISTRIBUTIONS

The Funnel, Squiggle and Rosenbrock distributions are smooth bijective transformations from a $Z \sim \mathcal{N}(\mu, \Sigma)$ to $X = f(X)$. We use the shorthand notation $x = x(z)$ and $z = z(x)$.

The Funnel distribution $p(x) = \mathcal{N}(x_D|0, \sigma^2)\mathcal{N}(x_{1:D-1}|\mu, e^{x_D} I_{D-1})$. In this case $Z \sim \mathcal{N}(0, I)$. The choice of parameters is $\sigma = 3$ and $\mu = 0$,

$$x = \begin{bmatrix} e^{\sigma z_D/2} z_{1:D-1} \\ \sigma z_D \end{bmatrix}, \quad \frac{\partial x}{\partial z} = \begin{bmatrix} e^{\sigma z_D/2} I_{D-1} & \frac{\sigma}{2} e^{\sigma z_D/2} z_{1:D-1} \\ 0 & \sigma \end{bmatrix} \quad \frac{\partial z}{\partial x} = \begin{bmatrix} e^{-x_D/2} I_{D-1} & -\frac{1}{2} e^{-x_D/2} x_{1:D-1} \\ 0 & \frac{1}{\sigma}. \end{bmatrix}.$$

The log determinant of the inverse Jacobian is $\log\det\left(\frac{\partial z}{\partial x}\right) = -(D-1)x_D/2 - \log\sigma$.

The hybrid Rosenbrock distribution For simplicity here we show the two dimensional case, the full distribution can be consulted in Pagani et al. [2022]. The two dimensional density is: $p(x) = \mathcal{N}(x_1|a, \frac{1}{2})\mathcal{N}(x_2|x_1^2, \frac{1}{2b})$. In this case $Z \sim \mathcal{N}(0, I)$. The choice of parameters is $a = 1$, $b = 100$ and block size of 3 and $\lfloor \frac{D-1}{3} \rfloor$ total blocks,

$$x = \begin{bmatrix} a + \frac{1}{\sqrt{2}} z_1 \\ (a + \frac{1}{\sqrt{2}} z_1)^2 + \frac{1}{\sqrt{2b}} z_2 \end{bmatrix}, \quad \frac{\partial x}{\partial z} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ \sqrt{2}a + z_1 & \frac{1}{\sqrt{2b}} \end{bmatrix}, \quad \frac{\partial z}{\partial x} = \begin{bmatrix} \sqrt{2} & 0 \\ -2\sqrt{2b} x_1 & \sqrt{2b} \end{bmatrix}.$$

The Squiggle distribution The density is $p(x) = \mathcal{N}(x(z)|\mu, \Sigma)|\det \frac{\partial x}{\partial z}|$, where $Z \sim \mathcal{N}(\mu, \Sigma)$. The choice of parameters is $a = 1.5$, $\mu = 0$, $\Sigma = \mathrm{diag}(5, \frac{1}{2}, .., \frac{1}{2})$

$$x = \begin{bmatrix} z_1 \\ z_{2:D} - \sin(a z_1) \end{bmatrix} \quad \frac{\partial x}{\partial z} = \begin{bmatrix} 1 & 0 \\ -a\cos(a z_1) & I \end{bmatrix}, \quad \frac{\partial z}{\partial x} = \begin{bmatrix} 1 & 0 \\ a\cos(a x_1) & I \end{bmatrix}.$$

The log determinant of the inverse Jacobian is $\log\det\left(\frac{\partial z}{\partial x}\right) = 0$.

For these three toy problems the Fisher Information follows from the transformation rule of Riemannian metrics

$$G(x) = \frac{\partial z}{\partial x}^\top \Sigma^{-1} \frac{\partial z}{\partial x}.$$

Location and Scale parameters for Complex Distributions In Experiment 5.3 we consider the mixture of two complex distributions. We introduce a location and scale parameter for each components of the mixture. The Funnel, Squiggle and Rosenbrock distributions are smooth bijective transformations from a $Z \sim \mathcal{N}(\mu, \Sigma)$ to $Y = f(X)$. Let us add a location and scale parameters by an additional transformation $g(Y) = X$, where $g(y) = \Sigma_y y + \mu_y$

$$Z \xrightarrow{f} Y \xrightarrow{g} X.$$

The change of variable formula for the composition $g \circ f$ gives

$$\begin{aligned} p_X(x) &= p_Z\left((g \circ f)^{-1}(x)\right)\left|\det \frac{\partial x}{\partial z}\right| \\ &= p_Z\left((f^{-1} \circ g^{-1})(x)\right)\left|\det \frac{\partial x}{\partial y}\right|\left|\det \frac{\partial y}{\partial z}\right| \end{aligned}$$

Plug in $g^{-1}(x) = \Sigma_y^{-1/2}(x - \mu_y)$, $\det \frac{\partial y}{\partial z} = \det \Sigma^{-1/2}$, and $p_Z(z) = \mathcal{N}(z|\mu, \Sigma)$, we obtain the expression of the density

$$p_X(x) = \mathcal{N}\left(f^{-1}(\Sigma^{1/2}(x - \mu_y))\middle|\mu, \Sigma\right)\left|\det \frac{\partial x}{\partial y}\right|\left|\det \Sigma_y^{-1/2}\right|.$$

Where (μ_y, Σ_y) are the location and scale parameters of the component of the mixture distribution.

The Allen-Cahn Field System We consider the stochastic Allen–Cahn model [Berglund et al., 2017] used as a benchmark in Cabezas et al. [2024b]. The log-density is:

$$\log p(x) = -\beta\left(\frac{a}{2\Delta s}\sum_{i=1}^{D+1}(x_i - x_{i-1})^2 + \frac{b\Delta s}{4}\sum_{i=1}^{D}(1 - x_i^2)^2\right). \tag{4}$$

We adopt the parameter choices $\Delta s = \frac{1}{D}$ and boundary conditions $x_0 = x_{D+1} = 0$, and the constants $a = 0.1$ and $b = \frac{1}{a}$ ensure that the double-well potential induces bimodality in each component x_i, and we fix $D = 16$.

Analysis of multimodality To understand the maxima of this density, we analyze the two terms in the log-density function (Eq. 4):

1. The first term, $\sum_{i=1}^{D+1}(x_i - x_{i-1})^2$, penalizes differences between adjacent components, encouraging all components to have similar values.
2. The second term, $\sum_{i=1}^{D}(1 - x_i^2)^2$, is minimized when $x_i = \pm 1$.

The global maxima occur at $(1, \ldots, 1)$ and $(-1, \ldots, -1)$ because these configurations minimize both terms simultaneously: all components have the same value (satisfying the first term) and each component equals ± 1 (satisfying the second term).

Local maxima occur at all other combinations of ± 1 values (i.e., at points $(\pm 1, \ldots, \pm 1)$ with mixed signs) because these configurations still satisfy the second term perfectly, but incur penalties from the first term due to sign changes between adjacent components.

This creates 2^D local maxima, making the problem highly multimodal, with the two homogeneous configurations being global maxima.

Kernel Stein Discrepancy Let π and ν be two probability measures. We estimate the Kernel Stein Discrepancy with the biased but non-negative V-estimator. Given a sample $x_i \sim \nu$ for $i = 1, \ldots, n$,

$$\widehat{\mathrm{KSD}}^2_{k,V}(\pi, \nu) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k_\pi(x_i, x_j').$$

Denote by $p(x)$ be the density of the measure π, then,

$$k_\pi(x, x') = \nabla_x \cdot \nabla_{x'} k(x, x') + \nabla_x k(x, x') \cdot \nabla_{x'} \log p(x') + \nabla_{x'} k(x, x') \cdot \nabla_x \log p(x) + k(x, x') \nabla_x \log p(x) \cdot \nabla_{x'} \log p(x),$$

where we choose the inverse multi quadratic kernel $k(x, x') = (1 + (x - x')^\top (x - x'))^\beta$ for $\beta = -\frac{1}{2}$, following the choices made by Cabezas et al. [2024b].