UC Irvine
UC Irvine Previously Published Works

Title
Sampling constrained probability distributions using Spherical Augmentation

Permalink
https://escholarship.org/uc/item/8v20w7nh

Authors
Lan, Shiwei
Shahbaba, Babak

Publication Date
2015-06-19

Peer reviewed

Sampling constrained probability distributions using Spherical Augmentation

Shiwei Lan   S.LAN@WARWICK.AC.UK
Department of Statistics
University of Warwick
Coventry CV4 7AL, UK

Babak Shahbaba   BABAKS@UCI.EDU
Department of Statistics and Department of Computer Science
University of California
Irvine, CA 92697, USA

Abstract

Statistical models with constrained probability distributions are abundant in machine learning. Some examples include regression models with norm constraints (e.g., Lasso), probit, many copula models, and latent Dirichlet allocation (LDA). Bayesian inference involving probability distributions confined to constrained domains could be quite challenging for commonly used sampling algorithms. In this paper, we propose a novel augmentation technique that handles a wide range of constraints by mapping the constrained domain to a sphere in the augmented space. By moving freely on the surface of this sphere, sampling algorithms handle constraints implicitly and generate proposals that remain within boundaries when mapped back to the original space. Our proposed method, called Spherical Augmentation, provides a mathematically natural and computationally efficient framework for sampling from constrained probability distributions. We show the advantages of our method over state-of-the-art sampling algorithms, such as exact Hamiltonian Monte Carlo, using several examples including truncated Gaussian distributions, Bayesian Lasso, Bayesian bridge regression, reconstruction of quantized stationary Gaussian process, and LDA for topic modeling.

Keywords: Constrained probability distribution; Geodesic; Hamiltonian; Monte Carlo; Lagrangian Monte Carlo

1. Introduction

Many commonly used statistical models in Bayesian analysis involve high-dimensional probability distributions confined to constrained domains. Some examples include regression models with norm constraints (e.g., Lasso), probit, many copula models, and latent Dirichlet allocation (LDA). Very often, the resulting models are intractable and simulating samples for Monte Carlo estimations is quite challenging (Neal and Roberts, 2008; Sherlock and Roberts, 2009; Neal et al., 2012; Brubaker et al., 2012; Pakman and Paninski, 2013). Although the literature on improving the efficiency of computational methods for Bayesian inference is quite extensive (see, for example, Neal, 1996, 1993; Geyer, 1992; Mykland et al., 1995; Propp and Wilson, 1996; Roberts and Sahu, 1997; Gilks et al., 1998; Warnes, 2001;

de Freitas et al., 2001; Brockwell, 2006; Neal, 2011, 2005, 2003; Beal, 2003; Møller et al., 2006; Andrieu and Moulines, 2006; Kurihara et al., 2006; Cappé et al., 2008; Craiu et al., 2009; Welling, 2009; Gelfand et al., 2010; Randal et al., 2007; Randal and P., 2011; Welling and Teh, 2011; Zhang and Sutton, 2011; Ahmadian et al., 2011; Girolami and Calderhead, 2011; Hoffman and Gelman, 2011; Beskos et al., 2011; Calderhead and Sustik, 2012; Shahbaba et al., 2014; Ahn et al., 2013; Lan et al., 2014; Ahn et al., 2014), these methods do not directly address the complications due to constrained target distributions. When dealing with such distributions, MCMC algorithms typically evaluate each proposal to ensure it is within the boundaries imposed by the constraints. Computationally, this is quite inefficient, especially in high dimensional problems where proposals are very likely to miss the constrained domain. Alternatively, one could map the original domain to the entire Euclidean space to remove the boundaries. This approach too is computationally inefficient since the sampler needs to explore a much larger space than needed.

In this paper, we propose a novel method, called *Spherical Augmentation*, for handling constraints involving norm inequalities (Figure 1). Our proposed method augments the parameter space and maps the constrained domain to a sphere in the augmented space. The sampling algorithm explores the surface of this sphere. This way, it handles constraints implicitly and generates proposals that remain within boundaries when mapped back to the original space. While our method can be applied to all Metropolis-based sampling algorithms, we mainly focus on methods based on Hamiltonian Monte Carlo (HMC) (Duane et al., 1987; Neal, 2011). As discussed by Neal (2011), one could modify standard HMC such that the sampler bounces off the boundaries by letting the potential energy go to infinity for parameter values that violate the constraints. This creates "energy walls" at boundaries. This approach, henceforth called *Wall HMC*, has limited applications and tends to be computationally inefficient, because the frequency of hitting and bouncing increases exponentially as dimension grows. Byrne and Girolami (2013) discuss an alternative approach for situations where constrained domains can be identified as sub-manifolds. Pakman and Paninski (2013) follow the idea of Wall HMC and propose an exact HMC algorithm specifically for truncated Gaussian distributions with non-holonomic constraints. Brubaker et al. (2012) on the other hand propose a modified version of HMC for handling holonomic constraint $c(\theta) = 0$. All these methods provide interesting solutions for specific types of constraints. In contrast, our proposed method offers a general and efficient framework applicable to a wide range of problems.

The paper is structured as follows. Before presenting our methods, in Section 2 we provide a brief overview of HMC and one of its variants, namely, Lagrangian Monte Carlo (LMC) (Lan et al., 2014). We then present the underlying idea of *spherical augmentation*, first for two simple cases, ball type (Section 3.1) and box type (Section 3.2) constraints, then for more general q-norm type constraints (Section 3.3), as well as some functional constraints (Section 3.4). In Section 4, we apply the spherical augmentation technique to HMC (Section 4.2) and LMC (Section 4.3) for sampling from constrained target distributions. We evaluate our proposed methods using simulated and real data in Section 5. Finally, Section 6 is devoted to discussion and future directions.
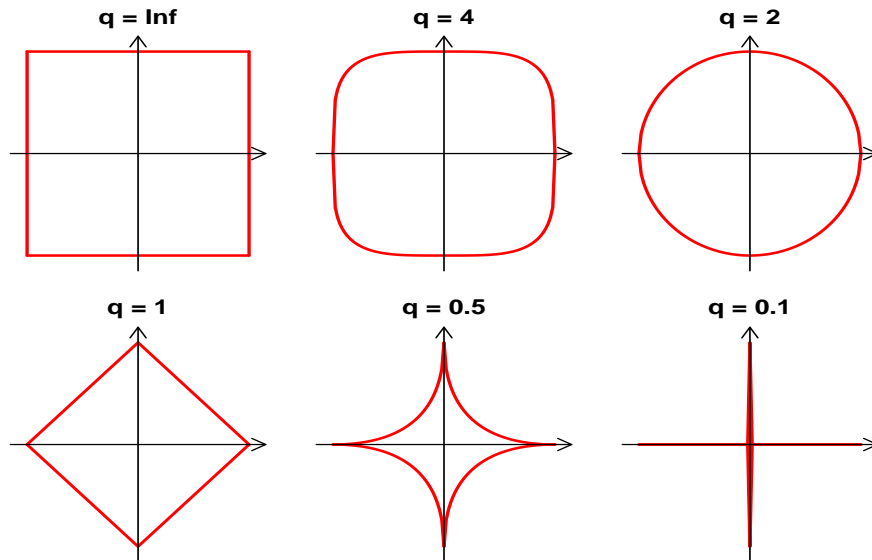
Figure 1: q-norm constraints

2. Preliminaries

2.1 Hamiltonian Monte Carlo

HMC improves upon random walk Metropolis (RWM) by proposing states that are distant from the current state, but nevertheless accepted with high probability. These distant proposals are found by numerically simulating Hamiltonian dynamics, whose state space consists of its *position*, denoted by the vector $\boldsymbol{\theta}$, and its *momentum*, denoted by the vector $\mathbf{p}$. Our objective is to sample from the continuous probability distribution of $\boldsymbol{\theta}$ with the density function $f(\boldsymbol{\theta})$. It is common to assume that the fictitious momentum variable $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{M})$, where $\mathbf{M}$ is a symmetric, positive-definite matrix known as the *mass matrix*, often set to the identity matrix $\mathbf{I}$ for convenience.

In this Hamiltonian dynamics, the *potential energy*, $U(\boldsymbol{\theta})$, is defined as minus the log density of $\boldsymbol{\theta}$ (plus any constant), that is $U(\boldsymbol{\theta}) := -\log f(\boldsymbol{\theta})$; the *kinetic energy*, $K(\mathbf{p})$ for the auxiliary momentum variable $\mathbf{p}$ is set to be minus the log density of $\mathbf{p}$ (plus any constant). Then the total energy of the system, *Hamiltonian* function, is defined as their sum,

$$H(\boldsymbol{\theta}, \mathbf{p}) = U(\boldsymbol{\theta}) + K(\mathbf{p}) \tag{1}$$

Given the Hamiltonian $H(\boldsymbol{\theta}, \mathbf{p})$, the system of $(\boldsymbol{\theta}, \mathbf{p})$ evolves according to the following *Hamilton's equations*,

$$\begin{aligned} \dot{\boldsymbol{\theta}} &= \nabla_{\mathbf{p}} H(\boldsymbol{\theta}, \mathbf{p}) &= \mathbf{M}^{-1} \mathbf{p} \\ \dot{\mathbf{p}} &= -\nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}, \mathbf{p}) &= -\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) \end{aligned} \tag{2}$$

In practice when the analytical solution to Hamilton's equations is not available, we need to numerically solve these equations by discretizing them, using some small time step ε. For the sake of accuracy and stability, a numerical method called *leapfrog* is commonly

used to approximate the Hamilton's equations (Neal, 2011). We usually solve the system for L steps, with some step size, ε, to propose a new state in the Metropolis algorithm, and accept or reject it according to the Metropolis acceptance probability. (See Neal, 2011, for more discussions).

2.2 Lagrangian Monte Carlo

Although HMC explores the target distribution more efficiently than RWM, it does not fully exploit its geometric properties of the parameter space. Girolami and Calderhead (2011) propose Riemannian HMC (RHMC), which adapts to the local Riemannian geometry of the target distribution by using a position-specific mass matrix $\mathbf{M} = \mathbf{G}(\boldsymbol{\theta})$. More specifically, they set $\mathbf{G}(\boldsymbol{\theta})$ to the Fisher information matrix. In this paper, we mainly use *spherical metric* instead to serve the purpose of constraint handling. The proposed method can be viewed as an extension to this approach since it explores the geometry of sphere.

Following the argument of Amari and Nagaoka (2000), Girolami and Calderhead (2011) define Hamiltonian dynamics on the Riemannian manifold endowed with metric $\mathbf{G}(\boldsymbol{\theta})$. With the non-flat metic, the momentum vector becomes $\mathbf{p}|\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}(\boldsymbol{\theta}))$ and the Hamiltonian is therefore defined as follows:

$$H(\boldsymbol{\theta}, \mathbf{p}) = \phi(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{p}^\mathsf{T}\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}, \quad \phi(\boldsymbol{\theta}) := U(\boldsymbol{\theta}) + \frac{1}{2}\log\det\mathbf{G}(\boldsymbol{\theta}) \tag{3}$$

Unfortunately the resulting Riemannian manifold Hamiltonian dynamics becomes non-separable since it contains products of $\boldsymbol{\theta}$ and $\mathbf{p}$, and the numerical integrator, *generalized leapfrog*, is an *implicit* scheme that involves time-consuming fixed-point iterations.

Lan et al. (2014) propose to change the variables $\mathbf{p} \mapsto \mathbf{v} := \mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}$ and define an *explicit* integrator for RHMC by using the following equivalent *Lagrangian* dynamics:

$$\dot{\boldsymbol{\theta}} = \mathbf{v} \tag{4}$$

$$\dot{\mathbf{v}} = -\mathbf{v}^\mathsf{T}\boldsymbol{\Gamma}(\boldsymbol{\theta})\mathbf{v} - \mathbf{G}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}) \tag{5}$$

where the *velocity* $\mathbf{v}|\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}(\boldsymbol{\theta})^{-1})$. Here, $\boldsymbol{\Gamma}(\boldsymbol{\theta})$ is the Christoffel Symbols derived from $\mathbf{G}(\boldsymbol{\theta})$.

The proposed *explicit* integrator is time reversible but not volume preserving. Based on the change of variables theorem, one can adjust the acceptance probability with Jacobian determinant to satisfy the detailed balance condition. The resulting algorithm, *Lagrangian Monte Carlo (LMC)*, is shown to be more efficient than RHMC (See Lan et al., 2014, for more details).

Throughout this paper, we express the kinetic energy K in terms of velocity, $\mathbf{v}$, instead of momentum, $\mathbf{p}$ (Beskos et al., 2011; Lan et al., 2014).

3. Spherical Augmentation

In this section, we introduce the *spherical augmentation* technique for handling norm constraints implicitly. We start with two simple constraints: ball type (2-norm) and box type (∞-norm). Then, we generalize the methodology to arbitrary q-norm type constraints for $q > 0$. Finally, we discuss some functional constraints that can be reduced to norm constraints.
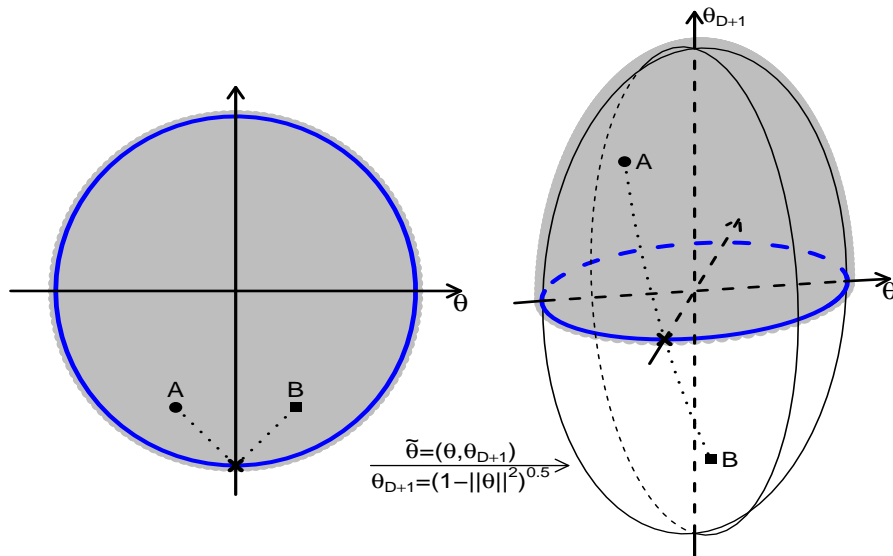
Figure 2: Transforming the unit ball $\mathcal{B}_0^D(1)$ to the sphere $\mathcal{S}^D$.

3.1 Ball type constraints

Consider probability distributions confined to the D-dimensional unit ball $\mathcal{B}_0^D(1) := \{\boldsymbol{\theta} \in \mathbb{R}^D : \|\boldsymbol{\theta}\|_2 = \sqrt{\sum_{i=1}^D \theta_i^2} \leq 1\}$. The constraint is given by restricting the 2-norm of parameters: $\|\boldsymbol{\theta}\|_2 \leq 1$.

The idea of spherical augmentation is to augment the original D-dimensional manifold of unit ball $\mathcal{B}_0^D(1)$ to a hyper-sphere $\mathcal{S}^D := \{\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{D+1} : \|\tilde{\boldsymbol{\theta}}\|_2 = 1\}$ in $(D+1)$-dimensional space. This can be done by adding an auxiliary variable θ_{D+1} to the original parameter $\boldsymbol{\theta} \in \mathcal{B}_0^D(1)$ to form an extended parameter $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}, \theta_{D+1})$ such that $\theta_{D+1} = \sqrt{1 - \|\boldsymbol{\theta}\|_2^2}$. Next, we identify the lower hemisphere $\mathcal{S}_-^D$ with the upper hemisphere $\mathcal{S}_+^D$ by ignoring the sign of θ_{D+1}. This way, the domain of the target distribution is changed from the unit ball $\mathcal{B}_0^D(1)$ to the D-dimensional sphere, $\mathcal{S}^D := \{\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{D+1} : \|\tilde{\boldsymbol{\theta}}\|_2 = 1\}$, through the following transformation:

$$T_{\mathcal{B} \to \mathcal{S}} : \mathcal{B}_0^D(1) \longrightarrow \mathcal{S}^D, \quad \boldsymbol{\theta} \mapsto \tilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}, \pm\sqrt{1 - \|\boldsymbol{\theta}\|_2^2}) \tag{6}$$

which can also be recognized as the coordinate map from the Euclidean coordinate chart $\{\boldsymbol{\theta}, \mathcal{B}_0^D(1)\}$ to the manifold $\mathcal{S}^D$.

After collecting samples $\{\tilde{\boldsymbol{\theta}}\}$ using a sampling algorithm (e.g., HMC) defined on the sphere, $\mathcal{S}^D$, we discard the last component θ_{D+1} and obtain the samples $\{\boldsymbol{\theta}\}$ that automatically satisfy the constraint $\|\boldsymbol{\theta}\|_2 \leq 1$. Note that the sign of θ_{D+1} does not affect our Monte Carlo estimates. However, after applying the above transformation, we need to adjust our estimates according to the change of variables theorem as follows:

$$\int_{\mathcal{B}_0^D(1)} f(\boldsymbol{\theta}) d\boldsymbol{\theta}_{\mathcal{B}} = \int_{\mathcal{S}_+^D} f(\tilde{\boldsymbol{\theta}}) \left| \frac{d\boldsymbol{\theta}_{\mathcal{B}}}{d\boldsymbol{\theta}_{\mathcal{S}_c}} \right| d\boldsymbol{\theta}_{\mathcal{S}_c} \tag{7}$$

where $\left| \frac{d\boldsymbol{\theta}_{\mathcal{B}}}{d\boldsymbol{\theta}_{\mathcal{S}_c}} \right| = |\theta_{D+1}|$ as shown in Corollary 1 in Appendix A.1. Here, $d\boldsymbol{\theta}_{\mathcal{B}}$ and $d\boldsymbol{\theta}_{\mathcal{S}_c}$ are volume elements under the Euclidean metric and the *canonical spherical metric* respectively.
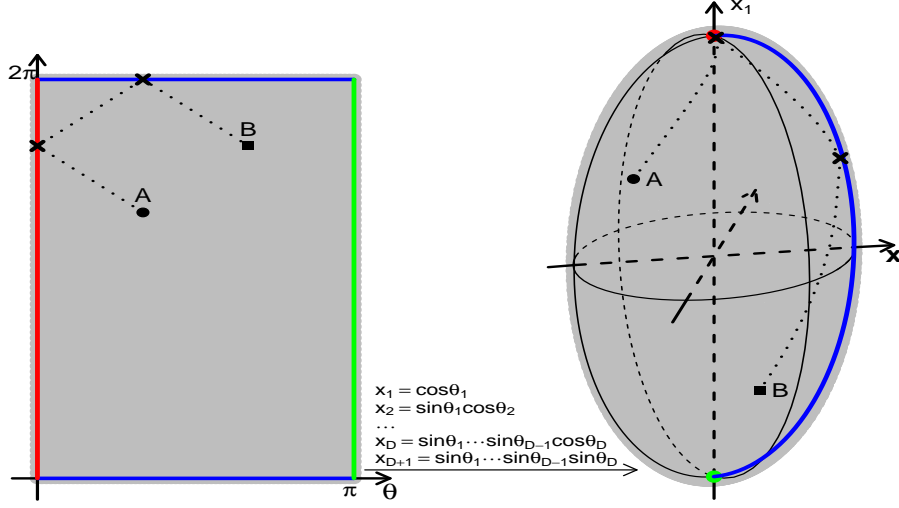
Figure 3: Transforming the hyper-rectangle $\mathcal{R}_0^D$ to the sphere $\mathcal{S}^D$.

With the above transformation (6), the resulting sampler is defined and moves freely on $\mathcal{S}^D$ while implicitly handling the constraints imposed on the original parameters. As illustrated in Figure 2, the boundary of the constraint, i.e., $\|\boldsymbol{\theta}\|_2 = 1$, corresponds to the equator on the sphere $\mathcal{S}^D$. Therefore, as the sampler moves on the sphere, e.g. from A to B, passing across the equator from one hemisphere to the other translates to "bouncing back" off the boundary in the original parameter space.

3.2 Box type constraints

Many constraints are given by both lower and upper bounds. Here we focus on a special case that defines a hyper-rectangle $\mathcal{R}_0^D := [0, \pi]^{D-1} \times [0, 2\pi)$; other *box* type constraints can be transformed to this hyper-rectangle. This constrained domain can be mapped to the unit ball $\mathcal{B}_0^D(1)$ and thus reduces to the ball type constraint discussed in Section 3.1. However, a more natural approach is to use *spherical* coordinates, which directly map the hyper-rectangle $\mathcal{R}_0^D$ to the sphere $\mathcal{S}^D$,

$$T_{\mathcal{R}_0 \to \mathcal{S}} : \mathcal{R}_0^D \longrightarrow \mathcal{S}^D, \quad \boldsymbol{\theta} \mapsto \mathbf{x}, \; x_d = \begin{cases} \cos(\theta_d) \prod_{i=1}^{d-1} \sin(\theta_i), & d < D+1 \\ \prod_{i=1}^{D} \sin(\theta_i), & d = D+1 \end{cases} \tag{8}$$

Therefore, we use $\{\boldsymbol{\theta}, \mathcal{R}_0^D\}$ as the spherical coordinate chart for the manifold $\mathcal{S}^D$. Instead of being appended with an extra dimension as in Section 3.1, here $\boldsymbol{\theta} \in \mathbb{R}^D$ is treated as the spherical coordinates of the point $\mathbf{x} \in \mathbb{R}^{D+1}$ with $\|\mathbf{x}\|_2 = 1$.

After obtaining samples $\{\mathbf{x}\}$ on the sphere $\mathcal{S}^D$, we transform them back to $\{\boldsymbol{\theta}\}$ in the original constrained domain $\mathcal{R}_0^D$ using the following inverse mapping of (8):

$$T_{\mathcal{S} \to \mathcal{R}_0} : \mathcal{S}^D \longrightarrow \mathcal{R}_0^D, \; \mathbf{x} \mapsto \boldsymbol{\theta}, \; \theta_d = \begin{cases} \operatorname{arccot} \dfrac{x_d}{\sqrt{1 - \sum_{i=1}^{d} x_i^2}}, & d < D \\ \operatorname{arccot} \dfrac{x_D}{x_{D+1}} + \dfrac{\pi}{2} \operatorname{sign}(x_{D+1})(\operatorname{sign}(x_{D+1}) - 1), & d = D \end{cases} \tag{9}$$
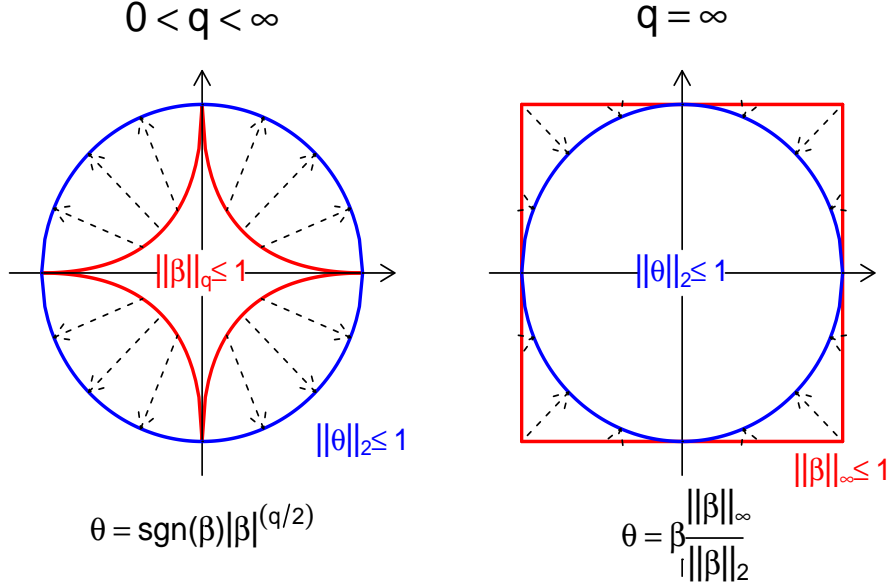
Figure 4: Transforming q-norm constrained domain to unit ball. Left: from unit cube $\mathcal{C}^D$ to unit ball $\mathcal{B}_0^D(1)$; Right from general q-norm domain $\mathcal{Q}^D$ to unit ball $\mathcal{B}_0^D(1)$.

Similarly, we need to adjust the estimates based on the following change of variables formula:

$$\int_{\mathcal{R}_0^D} f(\boldsymbol{\theta}) d\boldsymbol{\theta}_{\mathcal{R}_0} = \int_{\mathcal{S}^D} f(\boldsymbol{\theta}) \left| \frac{d\boldsymbol{\theta}_{\mathcal{R}_0}}{d\boldsymbol{\theta}_{\mathcal{S}_r}} \right| d\boldsymbol{\theta}_{\mathcal{S}_r} \tag{10}$$

where $\left| \frac{d\boldsymbol{\theta}_{\mathcal{R}_0}}{d\boldsymbol{\theta}_{\mathcal{S}_r}} \right| = \prod_{d=1}^{D-1} \sin^{-(D-d)}(\theta_d)$ as shown Proposition A.3 in Appendix A.3. Here, $d\boldsymbol{\theta}_{\mathcal{R}_0}$ and $d\boldsymbol{\theta}_{\mathcal{S}_r}$ are volume elements under the Euclidean metric and the *round spherical metric* respectively.

With the above transformation (8), we can derive sampling methods on the sphere to implicitly handle box type constraints. As illustrated in Figure 3, the red vertical boundary of $\mathcal{R}_0^D$ collapses to the north pole of $\mathcal{S}^D$, while the green vertical boundary collapses to the south pole. Two blue horizontal boundaries are mapped to the same prime meridian of $\mathcal{S}^D$ shown in blue color. As the sampler moves freely on the sphere $\mathcal{S}^D$, the resulting samples automatically satisfy the original constraint after being transformed back to the original domain.

3.3 General q-norm constraints

The ball and box type constraints discussed in previous sections are in fact special cases of more general q-norm constraints with q set to 2 and ∞ respectively. In general, these constraints are expressed in terms of q-norm of the parameter vector $\boldsymbol{\beta} \in \mathbb{R}^D$,

$$\|\boldsymbol{\beta}\|_q = \begin{cases} (\sum_{i=1}^D |\beta_i|^q)^{1/q}, & q \in (0, +\infty) \\ \max_{1 \le i \le D} |\beta_i|, & q = +\infty \end{cases} \tag{11}$$

7

This class of constraints is very common in statistics and machine learning. For example, when $\boldsymbol{\beta}$ are regression parameters, $q = 2$ corresponds to the ridge regression and $q = 1$ corresponds to Lasso (Tibshirani, 1996).

Denote the domain constrained by general q-norm as $\mathcal{Q}^D := \{\boldsymbol{\beta} \in \mathbb{R}^D : \|\boldsymbol{\beta}\|_q \le 1\}$. It could be quite challenging to sample probability distributions defined on $\mathcal{Q}^D$ (see Figure 1). To address this issue, we propose to transform $\mathcal{Q}^D$ to the unit ball $\mathcal{B}_{\mathbf{0}}^D(1)$ so that the method discussed in Section 3.1 can be applied. As before, sampling methods defined on the sphere $\mathcal{S}^D$ generate samples that automatically fall within $\mathcal{B}_{\mathbf{0}}^D(1)$. Then we transform those samples back to the q-norm domain, $\mathcal{Q}^D$, and adjust the estimates with the following change of variables formula:

$$\int_{\mathcal{Q}^D} f(\boldsymbol{\beta}) d\boldsymbol{\beta}_{\mathcal{Q}} = \int_{\mathcal{S}_+^D} f(\tilde{\boldsymbol{\theta}}) \left| \frac{d\boldsymbol{\beta}_{\mathcal{Q}}}{d\boldsymbol{\theta}_{\mathcal{S}_c}} \right| d\boldsymbol{\theta}_{\mathcal{S}_c} \tag{12}$$

where $\left| \frac{d\boldsymbol{\beta}_{\mathcal{Q}}}{d\boldsymbol{\theta}_{\mathcal{S}_c}} \right| = \left| \frac{d\boldsymbol{\beta}_{\mathcal{Q}}}{d\boldsymbol{\theta}_{\mathcal{B}}^{\mathsf{T}}} \right| \left| \frac{d\boldsymbol{\theta}_{\mathcal{B}}}{d\boldsymbol{\theta}_{\mathcal{S}_c}} \right| = \left| \frac{d\boldsymbol{\beta}_{\mathcal{Q}}}{d\boldsymbol{\theta}_{\mathcal{B}}^{\mathsf{T}}} \right| |\theta_{D+1}|$. In the following, we introduce the bijective mappings between $\mathcal{Q}^D$ and $\mathcal{B}_{\mathbf{0}}^D(1)$ and specify the associated Jacobian determinants $\left| \frac{d\boldsymbol{\beta}_{\mathcal{Q}}}{d\boldsymbol{\theta}_{\mathcal{B}}^{\mathsf{T}}} \right|$.

3.3.1 NORM CONSTRAINTS WITH $q \in (0, +\infty)$

For $q \in (0, +\infty)$, q-norm domain $\mathcal{Q}^D$ can be transformed to the unit ball $\mathcal{B}_0^D(1)$ bijectively via the following map (illustrated by the left panel of Figure 4):

$$T_{\mathcal{Q} \to \mathcal{B}} : \mathcal{Q}^D \to \mathcal{B}_0^D(1), \quad \beta_i \mapsto \theta_i = \mathrm{sgn}(\beta_i)|\beta_i|^{q/2} \tag{13}$$

The Jacobian determinant of $T_{\mathcal{B} \to \mathcal{Q}}$ is $\left| \frac{d\boldsymbol{\beta}_{\mathcal{Q}}}{d\boldsymbol{\theta}_{\mathcal{B}}^{\mathsf{T}}} \right| = \left(\frac{2}{q} \right)^D \left(\prod_{i=1}^D |\theta_i| \right)^{2/q-1}$. See Appendix B for more details.

3.3.2 NORM CONSTRAINTS WITH $q = +\infty$

When $q = +\infty$, the norm inequality defines a unit *hypercube*, $\mathcal{C}^D := [-1, 1]^D = \{\boldsymbol{\beta} \in \mathbb{R}^D : \|\boldsymbol{\beta}\|_\infty \le 1\}$, from which the more general form, *hyper-rectangle*, $\mathcal{R}^D := \{\boldsymbol{\beta} \in \mathbb{R}^D : \mathbf{l} \le \boldsymbol{\beta} \le \mathbf{u}\}$, can be obtained by proper shifting and scaling. The unit hypercube $\mathcal{C}^D$ can be transformed to its inscribed unit ball $\mathcal{B}_{\mathbf{0}}^D(1)$ through the following map (illustrated by the right panel of Figure 4):

$$T_{\mathcal{C} \to \mathcal{B}} : [-1, 1]^D \to \mathcal{B}_0^D(1), \quad \boldsymbol{\beta} \mapsto \boldsymbol{\theta} = \boldsymbol{\beta} \frac{\|\boldsymbol{\beta}\|_\infty}{\|\boldsymbol{\beta}\|_2} \tag{14}$$

The Jacobian determinant of $T_{\mathcal{B} \to \mathcal{R}}$ is $\left| \frac{d\boldsymbol{\beta}_{\mathcal{R}}}{d\boldsymbol{\theta}_{\mathcal{B}}^{\mathsf{T}}} \right| = \frac{\|\boldsymbol{\theta}\|_2^D}{\|\boldsymbol{\theta}\|_\infty^D} \prod_{i=1}^D \frac{u_i - l_i}{2}$. More details can be found in Appendix B.

3.4 Functional constraints

Many statistical problems involve functional constraints. For example, Pakman and Paninski (2013) discuss linear and quadratic constraints for multivariate Gaussian distributions. Since the target distribution is truncated Gaussian, Hamiltonian dynamics can be exactly
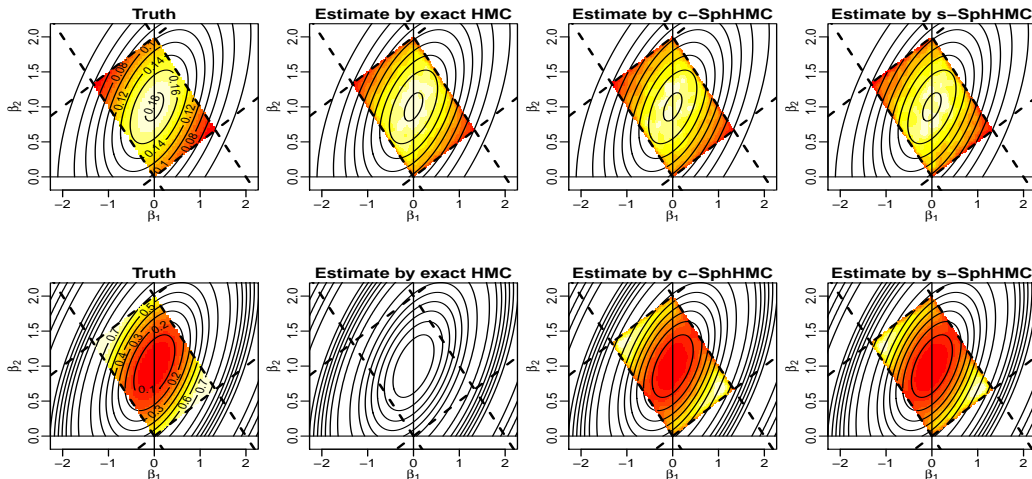
Figure 5: Sampling from a Gaussian distribution (first row) and a damped sine wave distribution (second row) with linear constraints. First column shows the true distributions. The exact HMC method of Pakman and Paninski (2013) is shown in the second column. The last two columns show our proposed methods.

simulated and the boundary-hitting time can be analytically obtained. However, finding the hitting time and reflection trajectory is computationally expensive. Some constraints of this type can be handled by the spherical augmentation method more efficiently. Further, our method can be use for sampling from a wide range of distributions beyond Gaussian.

3.4.1 Linear constraints

In general, M linear constraints can be written as $\mathbf{l} \leq \mathbf{A}\boldsymbol{\beta} \leq \mathbf{u}$, where $\mathbf{A}$ is $M \times D$ matrix, $\boldsymbol{\beta}$ is a D-vector, and the boundaries $\mathbf{l}$ and $\mathbf{u}$ are both M-vectors. Here, we assume $M = D$ and $\mathbf{A}_{D \times D}$ is invertible. (Note that we generally do not have $\mathbf{A}^{-1}\mathbf{l} \leq \boldsymbol{\beta} \leq \mathbf{A}^{-1}\mathbf{u}$.) Instead of sampling $\boldsymbol{\beta}$ directly, we sample $\boldsymbol{\eta} := \mathbf{A}\boldsymbol{\beta}$ with the box type constraint: $\mathbf{l} \leq \boldsymbol{\eta} \leq \mathbf{u}$. Now we can apply our proposed method to sample $\boldsymbol{\eta}$ and transform it back to $\boldsymbol{\beta} = \mathbf{A}^{-1}\boldsymbol{\eta}$. In this process, we use the following change of variables formula:

$$\int_{\mathbf{l} \leq \mathbf{A}\boldsymbol{\beta} \leq \mathbf{u}} f(\boldsymbol{\beta}) d\boldsymbol{\beta} = \int_{\mathbf{l} \leq \boldsymbol{\eta} \leq \mathbf{u}} f(\boldsymbol{\eta}) \left| \frac{d\boldsymbol{\beta}}{d\boldsymbol{\eta}} \right| d\boldsymbol{\eta} \qquad (15)$$

where $\left| \frac{d\boldsymbol{\beta}}{d\boldsymbol{\eta}} \right| = |\mathbf{A}|^{-1}$.

Figure 5 illustrates that both exact HMC (Pakman and Paninski, 2013) and HMC with spherical augmentation can handle linear constraints, here $\mathbf{l} = \mathbf{0}$, $\mathbf{A} = \begin{bmatrix} -0.5 & 1 \\ 1 & 1 \end{bmatrix}$ and $\mathbf{u} = \mathbf{2}$, imposed on a 2d Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ (first row). However, the exact HMC is not applicable to more complicated distributions such as

the damped sine wave distribution (second row in Figure 5) with the following density:

$$f(\boldsymbol{\beta}) \propto \frac{\sin^2 Q(\boldsymbol{\beta})}{Q(\boldsymbol{\beta})}, \quad Q(\boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}) \tag{16}$$

However, it is worth noting that for truncated Gaussian distributions, the exact HMC method of Pakman and Paninski (2013) can handle a wider range of linear constraints compared to our method.

3.4.2 QUADRATIC CONSTRAINTS

General quadratic constraints can be written as $l \le \boldsymbol{\beta}^\mathsf{T} \mathbf{A} \boldsymbol{\beta} + \mathbf{b}^\mathsf{T} \boldsymbol{\beta} \le u$, where $l, u > 0$ are scalars. We assume $\mathbf{A}_{D \times D}$ symmetric and positive definite. By spectrum theorem, we have the decomposition $\mathbf{A} = \mathbf{Q} \boldsymbol{\Sigma} \mathbf{Q}^\mathsf{T}$, where $\mathbf{Q}$ is an orthogonal matrix and $\boldsymbol{\Sigma}$ is a diagonal matrix of eigenvalues of $\mathbf{A}$. By shifting and scaling, $\boldsymbol{\beta} \mapsto \boldsymbol{\beta}^* = \sqrt{\boldsymbol{\Sigma}} \mathbf{Q}^\mathsf{T}(\boldsymbol{\beta} + \frac{1}{2}\mathbf{A}^{-1}\mathbf{b})$, we only need to consider the *ring* type constraints for $\boldsymbol{\beta}^*$,

$$\odot : l^* \le \|\boldsymbol{\beta}^*\|_2^2 = (\boldsymbol{\beta}^*)^\mathsf{T} \boldsymbol{\beta}^* \le u^*, \quad l^* = l + \frac{1}{4}\mathbf{b}^\mathsf{T} \mathbf{A}^{-1}\mathbf{b}, \; u^* = u + \frac{1}{4}\mathbf{b}^\mathsf{T} \mathbf{A}^{-1}\mathbf{b} \tag{17}$$

which can be mapped to the unit ball as follows:

$$T_{\odot \to \mathcal{B}} : \mathcal{B}_{\mathbf{0}}^D(\sqrt{u^*}) \backslash \mathcal{B}_{\mathbf{0}}^D(\sqrt{l^*}) \longrightarrow \mathcal{B}_{\mathbf{0}}^D(1), \quad \boldsymbol{\beta}^* \mapsto \boldsymbol{\theta} = \frac{\boldsymbol{\beta}^*}{\|\boldsymbol{\beta}^*\|_2} \frac{\|\boldsymbol{\beta}^*\|_2 - \sqrt{l^*}}{\sqrt{u^*} - \sqrt{l^*}} \tag{18}$$

We can then apply our proposed method in Section 3.1 to obtain samples $\{\boldsymbol{\theta}\}$ in $\mathcal{B}_{\mathbf{0}}^D(1)$ and transform them back to the original domain with the following inverse operation of (18):

$$T_{\mathcal{B} \to \odot} : \mathcal{B}_{\mathbf{0}}^D(1) \longrightarrow \mathcal{B}_{\mathbf{0}}^D(\sqrt{u^*}) \backslash \mathcal{B}_{\mathbf{0}}^D(\sqrt{l^*}), \quad \boldsymbol{\theta} \mapsto \boldsymbol{\beta}^* = \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_2}((\sqrt{u^*} - \sqrt{l^*})\|\boldsymbol{\theta}\|_2 + \sqrt{l^*}) \tag{19}$$

In this process, we need the change of variables formula

$$\int_{l \le \boldsymbol{\beta}^\mathsf{T} \mathbf{A} \boldsymbol{\beta} + \mathbf{b}^\mathsf{T} \boldsymbol{\beta} \le u} f(\boldsymbol{\beta})d\boldsymbol{\beta} = \int_{\mathcal{S}_+^D} f(\boldsymbol{\theta}) \left| \frac{d\boldsymbol{\beta}}{d\boldsymbol{\theta}_{\mathcal{S}_c}} \right| d\boldsymbol{\theta}_{\mathcal{S}_c} \tag{20}$$

where $\left| \frac{d\boldsymbol{\beta}}{d\boldsymbol{\theta}_{\mathcal{S}_c}} \right| = \left| \frac{d\boldsymbol{\beta}}{d(\boldsymbol{\beta}^*)^\mathsf{T}} \right| \left| \frac{d\boldsymbol{\beta}^*}{d\boldsymbol{\theta}_{\mathcal{B}}^\mathsf{T}} \right| \left| \frac{d\boldsymbol{\theta}_{\mathcal{B}}}{d\boldsymbol{\theta}_{\mathcal{S}_c}} \right| = |\mathbf{A}|^{-\frac{1}{2}} \alpha^{D-1}(\alpha - \sqrt{l^*})|\theta_{D+1}|, \; \alpha = \sqrt{u^*} + (1/\|\boldsymbol{\theta}\|_2 - 1)\sqrt{l^*}$.

3.4.3 MORE GENERAL CONSTRAINTS

We close this section with some comments on more general types of constraints. In some problems, several parameters might be unconstrained, and the type of constraints might be vary across the constrained parameters. In such cases, we could group the parameters into blocks and update each block separately using the methods discussed in this section. When dealing with one-sided constraints, e.g. $\theta_i \ge l_i$, one can map the constrained domain to the whole space and sample the unconstrained parameter θ_i^*, where $\theta_i = |\theta_i^*| + l_i$. Alternatively, the one-sided constraint $\theta_i \ge l_i$ can be changed to a two-sided constraint for $\theta_i^* \in (0, 1)$ by setting $\theta_i = -\log \theta_i^* + l_i$.

4. Monte Carlos with Spherical Augmentation

In this section, we show how the idea of spherical augmentation can be used to improve Markov Chain Monte Carlo methods applied to constrained probability distributions. In particular, we focus on two state-of-the-art sampling algorithms, namely Hamiltonian Monte Carlo(Duane et al., 1987; Neal, 2011), and Lagrangian Monte Carlo(Lan et al., 2014). Note however that our proposed method is generic so its application goes beyond these two algorithms.

4.1 Common settings

Throughout this section, we denote the original parameter vector as $\boldsymbol{\beta}$, the constrained domain as $\mathcal{D}$, the coordinate vector of sphere $\mathcal{S}^D$ as $\boldsymbol{\theta}$. All the change of variables formulae presented in the previous section can be summarized as

$$\int_{\mathcal{D}} f(\boldsymbol{\beta}) d\boldsymbol{\beta}_{\mathcal{D}} = \int_{\mathcal{S}} f(\boldsymbol{\theta}) \left| \frac{d\boldsymbol{\beta}_{\mathcal{D}}}{d\boldsymbol{\theta}_{\mathcal{S}}} \right| d\boldsymbol{\theta}_{\mathcal{S}} \tag{21}$$

where $\left| \frac{d\boldsymbol{\beta}_{\mathcal{D}}}{d\boldsymbol{\theta}_{\mathcal{S}}} \right|$ is the Jacobian determinant of the mapping $T : \mathcal{S} \longrightarrow \mathcal{D}$ and $d\boldsymbol{\theta}_{\mathcal{S}}$ is some spherical measure.

For energy based MCMC algorithms like HMC, RHMC and LMC, we need to investigate the change of energy under the above transformation. The original potential energy function $U(\boldsymbol{\beta}) = -\log f(\boldsymbol{\beta})$ should be transformed to the following $\phi(\boldsymbol{\theta})$

$$\phi(\boldsymbol{\theta}) = -\log f(\boldsymbol{\theta}) - \log \left| \frac{d\boldsymbol{\beta}_{\mathcal{D}}}{d\boldsymbol{\theta}_{\mathcal{S}}} \right| = U(\boldsymbol{\beta}(\boldsymbol{\theta})) - \log \left| \frac{d\boldsymbol{\beta}_{\mathcal{D}}}{d\boldsymbol{\theta}_{\mathcal{S}}} \right| \tag{22}$$

Consequently the total energy $H(\boldsymbol{\beta}, \mathbf{v})$ in (1) becomes

$$H(\boldsymbol{\theta}, \mathbf{v}) = \phi(\boldsymbol{\theta}) + \frac{1}{2} \langle \mathbf{v}, \mathbf{v} \rangle_{\mathbf{G}_{\mathcal{S}}(\boldsymbol{\theta})} \tag{23}$$

The gradient of potential energy U, metric and natural gradient (preconditioned gradient) under the new coordinate system $\{\boldsymbol{\theta}, \mathcal{S}^D\}$ can be calculated as follows

$$\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) = \frac{d\boldsymbol{\beta}^{\mathsf{T}}}{d\boldsymbol{\theta}} \nabla_{\boldsymbol{\beta}} U(\boldsymbol{\beta}) \tag{24}$$

$$\mathbf{G}_{\mathcal{S}}(\boldsymbol{\theta}) = \frac{d\boldsymbol{\beta}^{\mathsf{T}}}{d\boldsymbol{\theta}} \mathbf{G}_{\mathcal{D}}(\boldsymbol{\beta}) \frac{d\boldsymbol{\beta}}{d\boldsymbol{\theta}^{\mathsf{T}}} \tag{25}$$

$$\mathbf{G}_{\mathcal{S}}(\boldsymbol{\theta})^{-1} \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) = \left[\frac{d\boldsymbol{\beta}^{\mathsf{T}}}{d\boldsymbol{\theta}} \right]^{-1} \mathbf{G}_{\mathcal{D}}(\boldsymbol{\beta})^{-1} \nabla_{\boldsymbol{\beta}} U(\boldsymbol{\beta}) \tag{26}$$

4.2 Spherical Hamiltonian Monte Carlo

We define HMC on the sphere $\mathcal{S}^D$ in two different coordinate systems: the *Cartesian coordinate* and the *spherical coordinate*. The former is applied to ball type constraints or those that could be converted to ball type constraints; the later is more suited for box type constraints. Besides the merit of implicitly handling constraints, HMC on sphere can take advantage of the splitting technique (Beskos et al., 2011; Shahbaba et al., 2014; Byrne and Girolami, 2013) to further improve its computational efficiency.

4.2.1 SPHERICAL HMC IN THE CARTESIAN COORDINATE

We first consider HMC for the target distribution with density $f(\boldsymbol{\theta})$ defined on the unit ball $\mathcal{B}_0^D(1)$ endowed with the Euclidean metric $\mathbf{I}$. The potential energy is defined as $U(\boldsymbol{\theta}) := -\log f(\boldsymbol{\theta})$. Associated with the auxiliary variable $\mathbf{v}$ (i.e., velocity), we define the kinetic energy $K(\mathbf{v}) = \frac{1}{2}\mathbf{v}^\mathsf{T}\mathbf{I}\mathbf{v}$ for $\mathbf{v} \in T_{\boldsymbol{\theta}}\mathcal{B}_0^D(1)$, which is a D-dimensional vector sampled from the tangent space of $\mathcal{B}_0^D(1)$. Therefore, the Hamiltonian is defined on $\mathcal{B}_0^D(1)$ as

$$H(\boldsymbol{\theta}, \mathbf{v}) = U(\boldsymbol{\theta}) + K(\mathbf{v}) = U(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{v}^\mathsf{T}\mathbf{I}\mathbf{v} \tag{27}$$

Under the transformation $T_{\mathcal{B}\to\mathcal{S}}$ in (6), the above Hamiltonian (27) on $\mathcal{B}_0^D(1)$ will be changed to the follwing Hamiltonian $H(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{v}})$ on $\mathcal{S}^D$ as in (23):

$$H(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{v}}) = \phi(\tilde{\boldsymbol{\theta}}) + \frac{1}{2}\mathbf{v}^\mathsf{T}\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})\mathbf{v} = U(\tilde{\boldsymbol{\theta}}) - \log\left|\frac{d\boldsymbol{\theta}_{\mathcal{B}}}{d\boldsymbol{\theta}_{\mathcal{S}_c}}\right| + \frac{1}{2}\mathbf{v}^\mathsf{T}\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})\mathbf{v} \tag{28}$$

where the potential energy $U(\tilde{\boldsymbol{\theta}}) = U(\boldsymbol{\theta})$ (i.e., the distribution is fully defined in terms of the original parameter $\boldsymbol{\theta}$, which are the first D elements of $\tilde{\boldsymbol{\theta}}$), and $\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta}) = \mathbf{I}_D + \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}/(1-\|\boldsymbol{\theta}\|_2^2)$ is the *canonical spherical metric*.

Viewing $\{\boldsymbol{\theta}, \mathcal{B}_0^D(1)\}$ as the Euclidean coordinate chart of manifold $(\mathcal{S}^D, \mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta}))$, we have the logorithm of volume adjustment, $\log\left|\frac{d\boldsymbol{\theta}_{\mathcal{B}}}{d\boldsymbol{\theta}_{\mathcal{S}_c}}\right| = -\frac{1}{2}\log|\mathbf{G}_{\mathcal{S}_c}| = \log|\theta_{D+1}|$ (See Appendix A.1). The last two terms in Equation (28) is the minus log density of $\mathbf{v}|\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})^{-1})$ (See Girolami and Calderhead, 2011; Lan et al., 2014, for more details). However, the derivative of log volume adjustment, θ_{D+1}^{-1}, contributes an extremely large component to the gradient of energy around the equator ($\theta_{D+1} = 0$), which in turn increases the numerical error in the discretized Hamiltonian dynamics. For the purpose of numerical stability, we instead consider the following *partial* Hamiltonian $H^*(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{v}})$ and leave the volume adjustment as weights to adjust the estimation of integration (21):

$$H^*(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{v}}) = U(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{v}^\mathsf{T}\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})\mathbf{v} \tag{29}$$

If we extend the velocity as $\tilde{\mathbf{v}} = (\mathbf{v}, v_{D+1})$ with $v_{D+1} = -\boldsymbol{\theta}^\mathsf{T}\mathbf{v}/\theta_{D+1}$, then $\tilde{\mathbf{v}}$ falls in the tangent space of the sphere, $T_{\tilde{\boldsymbol{\theta}}}\mathcal{S}^D := \{\tilde{\mathbf{v}} \in \mathbb{R}^{D+1} | \tilde{\boldsymbol{\theta}}^\mathsf{T}\tilde{\mathbf{v}} = 0\}$. Therefore, $\mathbf{v}^\mathsf{T}\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})\mathbf{v} = \tilde{\mathbf{v}}^\mathsf{T}\tilde{\mathbf{v}}$. As a result, the partial Hamiltonian (29) can be recognized as the standard Hamiltonian (27) in the augmented $(D+1)$ dimensional space

$$H^*(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{v}}) = U(\tilde{\boldsymbol{\theta}}) + K(\tilde{\mathbf{v}}) = U(\tilde{\boldsymbol{\theta}}) + \frac{1}{2}\tilde{\mathbf{v}}^\mathsf{T}\tilde{\mathbf{v}} \tag{30}$$

This is due to the energy invariance presented as Proposition A.1 in Appendix A. Now we can sample the velocity $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})^{-1})$ and set $\tilde{\mathbf{v}} = \begin{bmatrix} \mathbf{I} \\ -\boldsymbol{\theta}^\mathsf{T}/\theta_{D+1} \end{bmatrix}\mathbf{v}$. Alternatively, since $\mathrm{Cov}[\tilde{\mathbf{v}}] = \begin{bmatrix} \mathbf{I} \\ -\boldsymbol{\theta}^\mathsf{T}/\theta_{D+1} \end{bmatrix}\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})^{-1}\begin{bmatrix} \mathbf{I} - \boldsymbol{\theta}/\theta_{D+1} \end{bmatrix} = \mathbf{I}_{D+1} - \tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}^\mathsf{T}$ is idempotent, we can sample $\tilde{\mathbf{v}}$ by $(\mathbf{I}_{D+1} - \tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}^\mathsf{T})\mathbf{z}$ with $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{D+1})$.

The Hamiltonian function (29) can be used to define the Hamiltonian dynamics on the Riemannian manifold $(\mathcal{S}^D, \mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta}))$ in terms of $(\boldsymbol{\theta}, \mathbf{p})$, or equivalently as the following Lagrangian dynamics in terms of $(\boldsymbol{\theta}, \mathbf{v})$ (Lan et al., 2014):

$$
\begin{aligned}
\dot{\boldsymbol{\theta}} &= \mathbf{v} \\
\dot{\mathbf{v}} &= -\mathbf{v}^{\mathsf{T}}\boldsymbol{\Gamma}_{\mathcal{S}_c}(\boldsymbol{\theta})\mathbf{v} - \mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta})
\end{aligned}
\tag{31}
$$

where $\boldsymbol{\Gamma}_{\mathcal{S}_c}(\boldsymbol{\theta})$ are the Christoffel symbols of second kind derived from $\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})$. The Hamiltonian (29) is preserved under Lagrangian dynamics (31). (See Lan et al., 2014, for more discussion).

Byrne and Girolami (2013) split the Hamiltonian (29) as follows:

$$
H^*(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{v}}) = U(\boldsymbol{\theta})/2 + \frac{1}{2}\mathbf{v}^{\mathsf{T}}\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})\mathbf{v} + U(\boldsymbol{\theta})/2
\tag{32}
$$

However, their approach requires the manifold to be embedded in the Euclidean space. To avoid this assumption, instead of splitting the Hamiltonian dynamics of $(\boldsymbol{\theta}, \mathbf{p})$, we split the corresponding Lagrangian dynamics (31) in terms of $(\boldsymbol{\theta}, \mathbf{v})$ as follows (See Appendix C for more details):

$$
\begin{cases}
\dot{\boldsymbol{\theta}} &= \mathbf{0} \\
\dot{\mathbf{v}} &= -\dfrac{1}{2}\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta})
\end{cases}
\tag{33a}
\qquad
\begin{cases}
\dot{\boldsymbol{\theta}} &= \mathbf{v} \\
\dot{\mathbf{v}} &= -\mathbf{v}^{\mathsf{T}}\boldsymbol{\Gamma}_{\mathcal{S}_c}(\boldsymbol{\theta})\mathbf{v}
\end{cases}
\tag{33b}
$$

Note that the first dynamics (33a) only involves updating velocity $\tilde{\mathbf{v}}$ in the tangent space $T_{\tilde{\boldsymbol{\theta}}}\mathcal{S}^D$ and has the following solution (see Appendix C for more details):

$$
\begin{aligned}
\tilde{\boldsymbol{\theta}}(t) &= \tilde{\boldsymbol{\theta}}(0) \\
\tilde{\mathbf{v}}(t) &= \tilde{\mathbf{v}}(0) - \frac{t}{2}\left(\begin{bmatrix}\mathbf{I}_D \\ \mathbf{0}^{\mathsf{T}}\end{bmatrix} - \tilde{\boldsymbol{\theta}}(0)\boldsymbol{\theta}(0)^{\mathsf{T}}\right)\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta}(0))
\end{aligned}
\tag{34}
$$

The second dynamics (33b) only involves the kinetic energy and has the geodesic flow that is a *great circle* (orthodrome or Riemannian circle) on the sphere $\mathcal{S}^D$ as its analytical solution (See Appendix A.2 for more details):

$$
\begin{aligned}
\tilde{\boldsymbol{\theta}}(t) &= \tilde{\boldsymbol{\theta}}(0)\cos(\|\tilde{\mathbf{v}}(0)\|_2 t) + \frac{\tilde{\mathbf{v}}(0)}{\|\tilde{\mathbf{v}}(0)\|_2}\sin(\|\tilde{\mathbf{v}}(0)\|_2 t) \\
\tilde{\mathbf{v}}(t) &= -\tilde{\boldsymbol{\theta}}(0)\|\tilde{\mathbf{v}}(0)\|_2\sin(\|\tilde{\mathbf{v}}(0)\|_2 t) + \tilde{\mathbf{v}}(0)\cos(\|\tilde{\mathbf{v}}(0)\|_2 t)
\end{aligned}
\tag{35}
$$

This solution defines an evolution, denoted as $g_t : (\boldsymbol{\theta}(0), \mathbf{v}(0)) \mapsto (\boldsymbol{\theta}(t), \mathbf{v}(t))$. Both (34) and (35) are symplectic. Due to the explicit formula for the geodesic flow on sphere, the second dynamics in (33b) is simulated exactly. Therefore, updating $\tilde{\boldsymbol{\theta}}$ does not involve discretization error so we can use large step sizes. This could lead to improved computational efficiency. Because this step is in fact a rotation on sphere, it can generate proposals that are far away from the current state. Algorithm 1 shows the steps for implementing this approach, henceforth called *Spherical HMC in the Cartesian coordinate (c-SphHMC)*. It can be shown that the integrator in the algorithm has order 3 local error and order 2 global error (See the details in Appendix D).

13

Algorithm 1 Spherical HMC in the Cartesian coordinate (c-SphHMC)

Initialize $\tilde{\boldsymbol{\theta}}^{(1)}$ at current $\tilde{\boldsymbol{\theta}}$ after transformation $T_{\mathcal{D}\to\mathcal{S}}$

Sample a new velocity value $\tilde{\mathbf{v}}^{(1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{D+1})$

Set $\tilde{\mathbf{v}}^{(1)} \leftarrow \tilde{\mathbf{v}}^{(1)} - \tilde{\boldsymbol{\theta}}^{(1)}(\tilde{\boldsymbol{\theta}}^{(1)})^{\mathsf{T}}\tilde{\mathbf{v}}^{(1)}$

Calculate $H(\tilde{\boldsymbol{\theta}}^{(1)}, \tilde{\mathbf{v}}^{(1)}) = U(\boldsymbol{\theta}^{(1)}) + K(\tilde{\mathbf{v}}^{(1)})$

for $\ell = 1$ to L **do**

$\quad \tilde{\mathbf{v}}^{(\ell+\frac{1}{2})} = \tilde{\mathbf{v}}^{(\ell)} - \frac{\varepsilon}{2}\left(\begin{bmatrix}\mathbf{I}_D \\ \mathbf{0}^{\mathsf{T}}\end{bmatrix} - \tilde{\boldsymbol{\theta}}^{(\ell)}(\boldsymbol{\theta}^{(\ell)})^{\mathsf{T}}\right)\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta}^{(\ell)})$

$\quad \tilde{\boldsymbol{\theta}}^{(\ell+1)} = \tilde{\boldsymbol{\theta}}^{(\ell)}\cos(\|\tilde{\mathbf{v}}^{(\ell+\frac{1}{2})}\|\varepsilon) + \frac{\tilde{\mathbf{v}}^{(\ell+\frac{1}{2})}}{\|\tilde{\mathbf{v}}^{(\ell+\frac{1}{2})}\|}\sin(\|\tilde{\mathbf{v}}^{(\ell+\frac{1}{2})}\|\varepsilon)$

$\quad \tilde{\mathbf{v}}^{(\ell+\frac{1}{2})} \leftarrow -\tilde{\boldsymbol{\theta}}^{(\ell)}\|\tilde{\mathbf{v}}^{(\ell+\frac{1}{2})}\|\sin(\|\tilde{\mathbf{v}}^{(\ell+\frac{1}{2})}\|\varepsilon) + \tilde{\mathbf{v}}^{(\ell+\frac{1}{2})}\cos(\|\tilde{\mathbf{v}}^{(\ell+\frac{1}{2})}\|\varepsilon)$

$\quad \tilde{\mathbf{v}}^{(\ell+1)} = \tilde{\mathbf{v}}^{(\ell+\frac{1}{2})} - \frac{\varepsilon}{2}\left(\begin{bmatrix}\mathbf{I}_D \\ \mathbf{0}^{\mathsf{T}}\end{bmatrix} - \tilde{\boldsymbol{\theta}}^{(\ell+1)}(\boldsymbol{\theta}^{(\ell+1)})^{\mathsf{T}}\right)\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta}^{(\ell+1)})$

end for

Calculate $H(\tilde{\boldsymbol{\theta}}^{(L+1)}, \tilde{\mathbf{v}}^{(L+1)}) = U(\boldsymbol{\theta}^{(L+1)}) + K(\tilde{\mathbf{v}}^{(L+1)})$

Calculate the acceptance probability $\alpha = \min\{1, \exp[-H(\tilde{\boldsymbol{\theta}}^{(L+1)}, \tilde{\mathbf{v}}^{(L+1)}) + H(\tilde{\boldsymbol{\theta}}^{(1)}, \tilde{\mathbf{v}}^{(1)})]\}$

Accept or reject the proposal according to α for the next state $\tilde{\boldsymbol{\theta}}'$

Calculate $T_{\mathcal{S}\to\mathcal{D}}(\tilde{\boldsymbol{\theta}}')$ and the corresponding weight $|dT_{\mathcal{S}\to\mathcal{D}}|$

4.2.2 Spherical HMC in the spherical coordinate

Now we define HMC on the sphere $\mathcal{S}^D$ in the spherical coordinate $\{\boldsymbol{\theta}, \mathcal{R}_\mathbf{0}^D\}$. The natural metric on the sphere $\mathcal{S}^D$ induced by the coordinate mapping (8) is the *round spherical metric*[1], $\mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta}) = \mathrm{diag}[1, \sin^2(\theta_1), \cdots, \prod_{d=1}^{D-1}\sin^2(\theta_d)]$.

As in Section 4.2.1, we start with the usual Hamiltonian $H(\boldsymbol{\theta}, \mathbf{v})$ defined on $(\mathcal{R}_\mathbf{0}^D, \mathbf{I})$ as in (27) with $\mathbf{v} \in T_{\boldsymbol{\theta}}\mathcal{R}_\mathbf{0}^D$. Under the transformation $T_{\mathcal{R}_\mathbf{0}\to\mathcal{S}} : \boldsymbol{\theta} \mapsto \mathbf{x}$ in (8), Hamiltonian (27) on $\mathcal{R}_\mathbf{0}^D$ is changed to the following Hamiltonian $H(\mathbf{x}, \dot{\mathbf{x}})$ as in (23):

$$H(\mathbf{x}, \dot{\mathbf{x}}) = \phi(\mathbf{x}) + \frac{1}{2}\mathbf{v}^{\mathsf{T}}\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})\mathbf{v} = U(\mathbf{x}) - \log\left|\frac{d\boldsymbol{\theta}_{\mathcal{R}_\mathbf{0}}}{d\boldsymbol{\theta}_{\mathcal{S}_r}}\right| + \frac{1}{2}\mathbf{v}^{\mathsf{T}}\mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta})\mathbf{v} \qquad (36)$$

where the potential energy $U(\mathbf{x}) = U(\boldsymbol{\theta}(\mathbf{x}))$ and $\mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta})$ is the round spherical metric.

As before, the logorithm of volume adjustment is $\log\left|\frac{d\boldsymbol{\theta}_{\mathcal{R}_\mathbf{0}}}{d\boldsymbol{\theta}_{\mathcal{S}_r}}\right| = -\frac{1}{2}\log|\mathbf{G}_{\mathcal{S}_r}| = -\sum_{d=1}^{D-1}(D-d)\log\sin(\theta_d)$ (See Appendix A.3). The last two terms in Equation (36) is the minus log density of $\mathbf{v}|\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta})^{-1})$. Again, for numerical stability we consider the following partial Hamiltonian $H^*(\mathbf{x}, \dot{\mathbf{x}})$ and leave the volume adjustment as weights to adjust the estimation of integration (21):

$$H^*(\mathbf{x}, \dot{\mathbf{x}}) = U(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{v}^{\mathsf{T}}\mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta})\mathbf{v} \qquad (37)$$

---

1. Note, $\mathbf{v}^{\mathsf{T}}\mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta})\mathbf{v} \le \|\mathbf{v}\|_2^2 \le \|\tilde{\mathbf{v}}\|_2^2 = \mathbf{v}^{\mathsf{T}}\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})\mathbf{v}$.

Taking derivative of $T_{\mathcal{R}_0 \to \mathcal{S}} : \boldsymbol{\theta} \mapsto \mathbf{x}$ in (8) with respect to time t we have

$$\dot{x}_d = \begin{cases} [-v_d \tan(\theta_d) + \sum_{i<d} v_i \cot(\theta_i)] x_d, & d < D+1 \\ \sum_{i<D+1} v_i \cot(\theta_i) x_{D+1}, & d = D+1 \end{cases} \tag{38}$$

We can show that $\mathbf{x}(\boldsymbol{\theta})^\mathsf{T} \dot{\mathbf{x}}(\boldsymbol{\theta}, \mathbf{v}) = 0$; that is, $\dot{\mathbf{x}} \in T_\mathbf{x} \mathcal{S}^D$. Taking derivative of $T_{\mathcal{S} \to \mathcal{R}_0} : \mathbf{x} \mapsto \boldsymbol{\theta}$ in (9) with respect to time t yields

$$v_d := \dot{\theta}_d = \begin{cases} -\dfrac{x_d}{\sqrt{1 - \sum_{i=1}^{d} x_i^2}} \left[\dfrac{\dot{x}_d}{x_d} + \dfrac{\sum_{i=1}^{d-1} x_i \dot{x}_i}{1 - \sum_{i=1}^{d-1} x_d^2} \right], & d < D \\ \dfrac{x_D \dot{x}_{D+1} - \dot{x}_D x_{D+1}}{x_D^2 + x_{D+1}^2}, & d = D \end{cases} \tag{39}$$

Further, we have $\mathbf{v}^\mathsf{T} \mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta}) \mathbf{v} = \dot{\mathbf{x}}^\mathsf{T} \dot{\mathbf{x}}$. Therefore, the partial Hamiltonian (37) can be recognized as the standard Hamiltonian (27) in the augmented $(D+1)$ dimensional space, which is again explained by the energy invariance Proposition A.1 (See more details in Appendix A)

$$H^*(\mathbf{x}, \dot{\mathbf{x}}) = U(\mathbf{x}) + K(\dot{\mathbf{x}}) = U(\mathbf{x}) + \frac{1}{2} \dot{\mathbf{x}}^\mathsf{T} \dot{\mathbf{x}} \tag{40}$$

Similar to the method discussed in Section 4.2.1, we split the Hamiltonian (37), $H^*(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{v}}) = U(\boldsymbol{\theta})/2 + \frac{1}{2} \mathbf{v}^\mathsf{T} \mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta}) \mathbf{v} + U(\boldsymbol{\theta})/2$, and its corresponding Lagrangian dynamics (31) as follows:

$$\begin{cases} \dot{\boldsymbol{\theta}} &= \mathbf{0} \\ \dot{\mathbf{v}} &= -\frac{1}{2} \mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta})^{-1} \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) \end{cases} \tag{41a} \qquad \begin{cases} \dot{\boldsymbol{\theta}} &= \mathbf{v} \\ \dot{\mathbf{v}} &= -\mathbf{v}^\mathsf{T} \boldsymbol{\Gamma}_{\mathcal{S}_r}(\boldsymbol{\theta}) \mathbf{v} \end{cases} \tag{41b}$$

The first dynamics (41a) involves updating the velocity $\mathbf{v}$ only. However, the diagonal term of $\mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta})^{-1}$, $\prod_{i=1}^{d-1} \sin^{-2}(\theta_i)$ increases exponentially fast as dimension grows. This will cause the velocity updated by (41a) to have extremely large components. To avoid such issue, we use small time vector $\boldsymbol{\varepsilon} = [\varepsilon, \varepsilon^2, \cdots, \varepsilon^D]$, instead of scalar ε, in updating Equation (41a). The second dynamics (41b) describes the same geodesic flow on the sphere $\mathcal{S}^D$ as (33b) but in the spherical coordinate $\{\boldsymbol{\theta}, \mathcal{R}_0^D\}$. Therefore it should have the same solution as (35) expressed in $\{\boldsymbol{\theta}, \mathcal{R}_0^D\}$. To obtain this solution, we first apply $\tilde{T}_{\mathcal{R}_0 \to \mathcal{S}} : (\boldsymbol{\theta}(0), \mathbf{v}(0)) \mapsto (\mathbf{x}(0), \dot{\mathbf{x}}(0))$, which consists of (8)(38). Then, we use g_t in (35) to evolve $(\mathbf{x}(0), \dot{\mathbf{x}}(0))$ for some time t to find $(\mathbf{x}(t), \dot{\mathbf{x}}(t))$. Finally, we use $\tilde{T}_{\mathcal{S} \to \mathcal{R}_0} : (\mathbf{x}(t), \dot{\mathbf{x}}(t)) \mapsto (\boldsymbol{\theta}(t), \mathbf{v}(t))$, composite of (9)(39), to go back to $\mathcal{R}_0^D$.

Algorithm 2 summarizes the steps for this method, called *Spherical HMC in the spherical coordinate (s-SphHMC)*. In theory, the hyper-rectangle $\mathcal{R}_0^D$ can be used as a base type (as the unit ball $\mathcal{B}_0^D(1)$ does) for general q-norm constraints for which s-SphHMC can be applied. This is because q-norm domain $\mathcal{Q}^D$ can be bijectively mapped to the hypercube $\mathcal{C}^D$, and thereafter to $\mathcal{R}_0^D$. However the involved Jacobian matrix is rather complicated and s-SphHMC used in this way is not as efficient as c-SphHMC. Therefore, we use s-SphHMC only for box type constraints.

Algorithm 2 Spherical HMC in the spherical coordinate (s-SphHMC)

Initialize $\boldsymbol{\theta}^{(1)}$ at current $\boldsymbol{\theta}$ after transformation $T_{\mathcal{D} \to \mathcal{S}}$

Sample a new velocity value $\mathbf{v}^{(1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$

Set $v_d^{(1)} \leftarrow v_d^{(1)} \prod_{i=1}^{d-1} \sin^{-1}(\theta_i^{(1)})$, $d = 1, \cdots, D$

Calculate $H(\boldsymbol{\theta}^{(1)}, \mathbf{v}^{(1)}) = U(\boldsymbol{\theta}^{(1)}) + K(\mathbf{v}^{(1)})$

for $\ell = 1$ to L **do**

$\qquad v_d^{(\ell+\frac{1}{2})} = v_d^{(\ell)} - \frac{\varepsilon^d}{2} \frac{\partial}{\partial \theta_d} U(\boldsymbol{\theta}^{(\ell)}) \prod_{i=1}^{d-1} \sin^{-2}(\theta_i^{(\ell)})$, $d = 1, \cdots, D$

$\qquad (\boldsymbol{\theta}^{(\ell+1)}, \mathbf{v}^{(\ell+\frac{1}{2})}) \leftarrow \tilde{T}_{\mathcal{S} \to \mathcal{R}_\mathbf{0}} \circ g_\varepsilon \circ \tilde{T}_{\mathcal{R}_\mathbf{0} \to \mathcal{S}}(\boldsymbol{\theta}^{(\ell)}, \mathbf{v}^{(\ell+\frac{1}{2})})$

$\qquad v_d^{(\ell+1)} = v_d^{(\ell+\frac{1}{2})} - \frac{\varepsilon^d}{2} \frac{\partial}{\partial \theta_d} U(\boldsymbol{\theta}^{(\ell+1)}) \prod_{i=1}^{d-1} \sin^{-2}(\theta_i^{(\ell+1)})$, $d = 1, \cdots, D$

end for

Calculate $H(\boldsymbol{\theta}^{(L+1)}, \mathbf{v}^{(L+1)}) = U(\boldsymbol{\theta}^{(L+1)}) + K(\mathbf{v}^{(L+1)})$

Calculate the acceptance probability $\alpha = \min\{1, \exp[-H(\boldsymbol{\theta}^{(L+1)}, \mathbf{v}^{(L+1)}) + H(\boldsymbol{\theta}^{(1)}, \mathbf{v}^{(1)})]\}$

Accept or reject the proposal according to α for the next state $\boldsymbol{\theta}'$

Calculate $T_{\mathcal{S} \to \mathcal{D}}(\boldsymbol{\theta}')$ and the corresponding weight $|dT_{\mathcal{S} \to \mathcal{D}}|$

4.3 Spherical LMC on probability simplex

A large class of statistical models involve defining probability distributions on the *simplex* Δ^K,

$$\Delta^K := \{\boldsymbol{\pi} \in \mathbb{R}^D | \ \pi_k \geq 0, \sum_{k=1}^{K} \pi_d = 1\} \tag{42}$$

As an example, we consider *latent Dirichlet allocation (LDA)* (Blei et al., 2003), which is a hierarchical Bayesian model commonly used to model document topics. This type of constraints can be viewed as a special case of the 1-norm constraint, discussed in Section 3.3.1, by identifying the first orthant (all positive components) with the others. Then, the c-SphHMC algorithm 1 can be applied to generate samples $\{\boldsymbol{\theta}\}$ on the sphere $\mathcal{S}^{K-1}$. These samples can be transformed as $\{\boldsymbol{\theta}^2\}$ and mapped back to the simplex Δ^K.

In what follows, we show that Fisher metric on the root space of simplex, $\sqrt{\Delta}^K := \{\boldsymbol{\theta} \in \mathcal{S}^{K-1} | \theta_k \geq 0, \forall k = 1, \cdots, K\}$ (i.e. the first orthant of the sphere $\mathcal{S}^{K-1}$), is the same as the canonical spherical metric $\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})$ up to a constant. In this sense, it is more natural to define the sampling algorithms on the sphere $\mathcal{S}^{K-1}$. We start with the toy example discussed in Patterson and Teh (2013). Denote the observed data as $\mathbf{x} = \{x_i\}_{i=1}^N$, where each data point belongs to one of the K categories with probability $p(x_i = k | \boldsymbol{\pi}) = \pi_k$. We assume a Dirichlet prior on $\boldsymbol{\pi}$: $p(\boldsymbol{\pi}) \propto \prod_{k=1}^K \pi_k^{\alpha_k - 1}$. The posterior distribution is $p(\boldsymbol{\pi} | \mathbf{x}) \propto \prod_{k=1}^K \pi_k^{n_k + \alpha_k - 1}$, where $n_k = \sum_{i=1}^N I(x_i = k)$ counts the points x_i in category k. Denote $\mathbf{n} = [n_1, \cdots, n_K]^\mathsf{T}$ and $n := |\mathbf{n}| = \sum_{k=1}^K n_k$. For inference, we need to sample from the posterior distribution $p(\boldsymbol{\pi} | \mathbf{x})$ defined on the probability simplex.
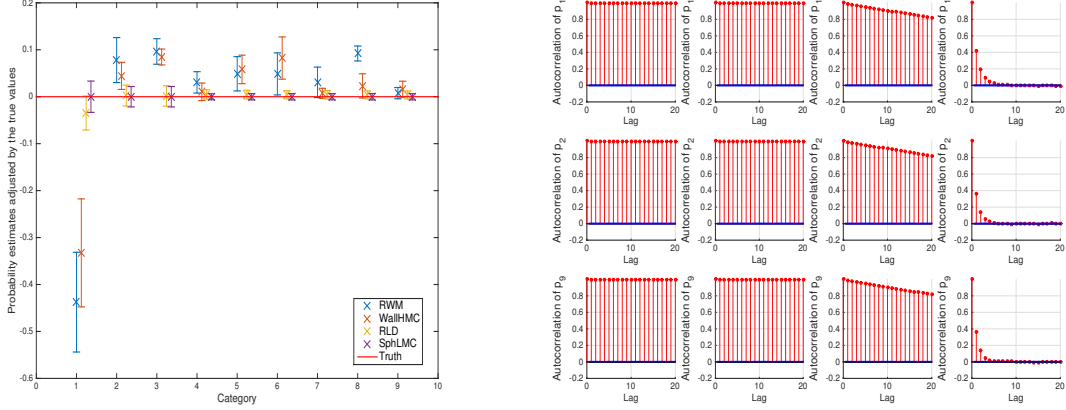
Figure 6: Dirichlet-Multinomial model: probability estimates (left) and autocorrelation function for MCMC samples (right).

The Fisher information matrix is a function of $\boldsymbol{\pi}_{-K}$ (here, '$-K$' means all but the K-th components) and is calculated as follows:

$$
\begin{aligned}
\mathbf{G_F}(\boldsymbol{\pi}_{-K}) &= -\mathrm{E}[\nabla^2 \log p(\mathbf{x}|\boldsymbol{\pi}_{-K})] \\
&= -\mathrm{E}[\nabla^2 (\mathbf{n}_{-K}^\mathsf{T} \log(\boldsymbol{\pi}_{-K}) + (n - \mathbf{n}_{-K}^\mathsf{T}\mathbf{1}) \log(1 - \boldsymbol{\pi}_{-K}^\mathsf{T}\mathbf{1}))] \\
&= -\mathrm{E}[\nabla(\mathbf{n}_{-K}/\boldsymbol{\pi}_{-K} - \mathbf{1}(n - \mathbf{n}_{-K}^\mathsf{T}\mathbf{1})/(1 - \boldsymbol{\pi}_{-K}^\mathsf{T}\mathbf{1}))] \\
&= -\mathrm{E}[-\operatorname{diag}(\mathbf{n}_{-K}/\boldsymbol{\pi}_{-K}^2) - \mathbf{1}\mathbf{1}^\mathsf{T}(n - \mathbf{n}_{-K}^\mathsf{T}\mathbf{1})/(1 - \boldsymbol{\pi}_{-K}^\mathsf{T}\mathbf{1})^2] \\
&= n[\operatorname{diag}(1/\boldsymbol{\pi}_{-K}) + \mathbf{1}\mathbf{1}^\mathsf{T}/\pi_K]
\end{aligned}
\tag{43}
$$

Now we use $T_{\Delta \to \sqrt{\Delta}} : \boldsymbol{\pi} \mapsto \boldsymbol{\theta} = \sqrt{\boldsymbol{\pi}}$ to map the simplex to the sphere (the first orthant). Note that $\frac{d\boldsymbol{\pi}_{-K}}{d\boldsymbol{\theta}_{-K}^\mathsf{T}} = 2\operatorname{diag}(\boldsymbol{\theta}_{-K})$. Therefore, we have a proper metric on $\sqrt{\Delta}^K$ as follows:

$$
\mathbf{G}_{\sqrt{\Delta}}(\boldsymbol{\theta}) = \frac{d\boldsymbol{\pi}_{-K}^\mathsf{T}}{d\boldsymbol{\theta}_{-K}} \mathbf{G_F}(\boldsymbol{\pi}_{-K}) \frac{d\boldsymbol{\pi}_{-K}}{d\boldsymbol{\theta}_{-K}^\mathsf{T}} = 4n[\mathbf{I}_{K-1} + \boldsymbol{\theta}_{-K}\boldsymbol{\theta}_{-K}^\mathsf{T}/\theta_K^2] = 4n\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})
\tag{44}
$$

where the scalar $4n$ properly scales the metric in high dimensional data intensive models. In LDA particularly, n could be the number of words counted in the selected documents. Hence, we use $\mathbf{G}_{\sqrt{\Delta}}(\boldsymbol{\theta})$ instead of $\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})$. We refer to the resulting method as *Spherical Lagrangian Monte Carlo (SphLMC)*.

Recall that in the development of Spherical HMC algorithms, we decided to omit the log volume adjustment term, $\log \left| \frac{d\boldsymbol{\beta}_{\mathcal{D}}}{d\boldsymbol{\theta}_{\mathcal{S}}} \right|$, in the partial Hamiltonian (29) and (37), and regard it as the weight to adjust the estimate of (21) or resample. This is not feasible if the LDA model is going to be used in an online setting. Therefore, we use $\phi(\boldsymbol{\theta})$ in (22), as opposed to $U(\boldsymbol{\theta})$ to avoid the re-weighting step.

17

To illustrate our proposed method, we consider the toy example discussed above. For this problem, Patterson and Teh (2013) propose a Riemannian Langevin Dynamics (RLD) method, but use an expanded-mean parametrization to map the simplex to the whole space. As mentioned above, this approach (i.e., expanding the parameter space) might not be efficient in general. This is illustrated in Figure 6. Here, we set $\alpha = 0.5$ and run RMW, WallHMC, RLD, and SphLMC[2] for 1.1×10^5 iterations; we discard the first 10^4 samples. As we can see in Figure 6, compared to alternative algorithms, our SphLMC method provides better probability estimates (left panel). Further, SphLMC generates samples with a substantially lower autocorrelation (right panel).

5. Experimental results

In this section, we evaluate our proposed methods using simulated and real data. To this end, we compare their efficiency to that of RWM, Wall HMC, exact HMC (Pakman and Paninski, 2013), and the Riemannian Langevin dynamics (RLD) algorithm proposed by Patterson and Teh (2013) for LDA. We define efficiency in terms of time-normalized effective sample size (ESS). Given N MCMC samples, for each parameter, we define ESS $= N[1+2\Sigma_{k=1}^{K}\rho(k)]^{-1}$, where $\rho(k)$ is sample autocorrelation with lag k (Geyer, 1992). We use the minimum ESS normalized by the CPU time, s (in seconds), as the overall measure of efficiency: min(ESS)/s. All computer codes are available online at `http://www.ics.uci.edu/~slan/SphHMC`.

5.1 Truncated Multivariate Gaussian

For illustration purpose, we start with a truncated bivariate Gaussian distribution,

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}\right), \qquad 0 \le \beta_1 \le 5, \quad 0 \le \beta_2 \le 1$$

This is box type constraint with the lower and upper limits as $\mathbf{l} = (0,0)$ and $\mathbf{u} = (5,1)$ respectively. The original rectangle domain can be mapped to 2d unit disc $\mathcal{B}_0^2(1)$ to use c-SphHMC, or mapped to 2d rectangle $\mathcal{R}_0^2$ where s-SphHMC can be directly applied.

The upper leftmost panel of Figure 7 shows the heatmap based on the exact density function, and the other panels show the corresponding heatmaps based on MCMC samples from RWM, Wall HMC, exact HMC, c-SphHMC and s-SphHMC respectively. Table 1 compares the true mean and covariance of the above truncated bivariate Gaussian distribution with the point estimates using 2×10^5 (2×10^4 for each of 10 repeated experiments with different random seeds) MCMC samples in each method. Overall, all methods estimate the mean and covariance reasonably well.

To evaluate the efficiency of the above-mentioned methods, we repeat this experiment for higher dimensions, $D = 10$, and $D = 100$. As before, we set the mean to zero and set the (i,j)-th element of the covariance matrix to $\Sigma_{ij} = 1/(1 + |i - j|)$. Further, we impose

---

2. Note, the natural gradient in (34) to update $\tilde{\mathbf{v}}$ is $\begin{bmatrix} \mathbf{I}_{K-1} \\ -\boldsymbol{\theta}_{-K}^{\mathsf{T}}/\theta_K \end{bmatrix} \mathbf{G}_{\sqrt{\Delta}}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}_{-K}}\phi(\boldsymbol{\theta}_{-K}) = [(\mathbf{n} + \alpha - 0.5)/\boldsymbol{\theta} - \boldsymbol{\theta} * |\mathbf{n} + \alpha - 0.5|]/(2n)$.
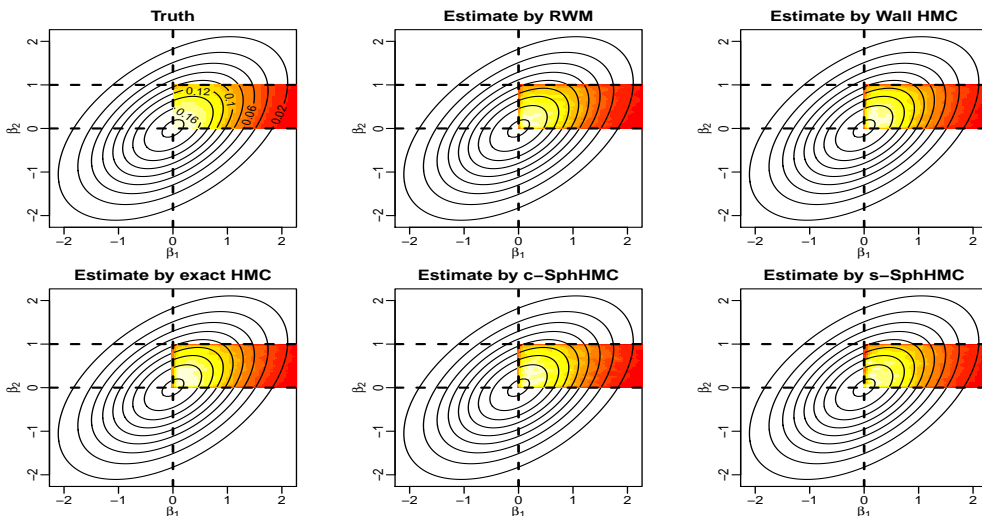
Figure 7: Density plots of a truncated bivariate Gaussian using exact density function (upper leftmost) and MCMC samples from RWM, Wall HMC, exact HMC, c-SphHMC and s-SphHMC respectively.

Method	Mean	Covariance	
Truth	$\begin{bmatrix} 0.7906 \\ 0.4889 \end{bmatrix}$	$\begin{bmatrix} 0.3269 & 0.0172 \\ 0.0172 & 0.08 \end{bmatrix}$	
RWM	$\begin{bmatrix} 0.7796 \pm 0.0088 \\ 0.4889 \pm 0.0034 \end{bmatrix}$	$\begin{bmatrix} 0.3214 \pm 0.009 & 0.0158 \pm 0.001 \\ 0.0158 \pm 0.001 & 0.0798 \pm 5e-04 \end{bmatrix}$	
Wall HMC	$\begin{bmatrix} 0.7875 \pm 0.0049 \\ 0.4884 \pm 8e-04 \end{bmatrix}$	$\begin{bmatrix} 0.3242 \pm 0.0043 & 0.017 \pm 0.001 \\ 0.017 \pm 0.001 & 0.08 \pm 3e-04 \end{bmatrix}$	
exact HMC	$\begin{bmatrix} 0.7909 \pm 0.0025 \\ 0.4885 \pm 0.001 \end{bmatrix}$	$\begin{bmatrix} 0.3272 \pm 0.0026 & 0.0174 \pm 7e-04 \\ 0.0174 \pm 7e-04 & 0.08 \pm 3e-04 \end{bmatrix}$	
c-SphHMC	$\begin{bmatrix} 0.79 \pm 0.005 \\ 0.4864 \pm 0.0016 \end{bmatrix}$	$\begin{bmatrix} 0.3249 \pm 0.0045 & 0.0172 \pm 0.0012 \\ 0.0172 \pm 0.0012 & 0.0801 \pm 0.001 \end{bmatrix}$	
s-SphHMC	$\begin{bmatrix} 0.7935 \pm 0.0093 \\ 0.4852 \pm 0.003 \end{bmatrix}$	$\begin{bmatrix} 0.3233 \pm 0.0062 & 0.0202 \pm 0.0018 \\ 0.0202 \pm 0.0018 & 0.0791 \pm 9e-04 \end{bmatrix}$	

Table 1: Comparing the point estimates for the mean and covariance of a bivariate truncated Gaussian distribution using RWM, Wall HMC, exact HMC, c-SphHMC and s-SphHMC.

the following constraints on the parameters,

$$0 \le \beta_i \le u_i$$

where u_i (i.e., the upper bound) is set to 5 when $i = 1$; otherwise, it is set to 0.5.

19

For each method, we obtain 10^5 MCMC samples after discarding the initial 10^4 samples. We set the tuning parameters of algorithms such that their overall acceptance rates are within a reasonable range. As shown in Table 2, Spherical HMC algorithms are substantially more efficient than RWM and Wall HMC. For RWM, the proposed states are rejected about 95% of times due to violation of the constraints. On average, Wall HMC bounces off the wall around 3.81 ($L = 2$) and 6.19 ($L = 5$) times per iteration for $D = 10$ and $D = 100$ respectively. Exact HMC is quite efficient for relatively low dimensional truncated Gaussian ($D = 10$); however it becomes very slow for higher dimensions ($D = 100$). In contrast, by augmenting the parameter space, Spherical HMC algorithms handle the constraints in a more efficient way. Since s-SphHMC is more suited for box type constraints, it is substantially more efficient than c-SphHMC in this example.

Dim	Method	AP	s/iter	ESS(min,med,max)	Min(ESS)/s	spdup
	RWM	0.62	5.72E-05	(48,691,736)	7.58	1.00
	Wall HMC	0.83	1.19E-04	(31904,86275,87311)	2441.72	322.33
D= 10	exact HMC	1.00	7.60E-05	(1e+05,1e+05,1e+05)	11960.29	1578.87
	c-SphHMC	0.82	2.53E-04	(62658,85570,86295)	2253.32	297.46
	s-SphHMC	0.79	2.02E-04	(76088,1e+05,1e+05)	3429.56	452.73
	RWM	0.81	5.45E-04	(1,4,54)	0.01	1.00
	Wall HMC	0.74	2.23E-03	(17777,52909,55713)	72.45	5130.21
D= 100	exact HMC	1.00	4.65E-02	(97963,1e+05,1e+05)	19.16	1356.64
	c-SphHMC	0.73	3.45E-03	(55667,68585,72850)	146.75	10390.94
	s-SphHMC	0.87	2.30E-03	(74476,99670,1e+05)	294.31	20839.43

Table 2: Comparing the efficiency of RWM, Wall HMC, exact HMC, c-SphHMC and s-SphHMC in terms of sampling from truncated Gaussian distributions. AP is acceptance probability, s/iter is seconds per iteration, ESS(min,med,max) is the (minimum, median, maximum) effective sample size, and Min(ESS)/s is the time-normalized minimum ESS.

5.2 Bayesian Lasso

In regression analysis, overly complex models tend to overfit the data. Regularized regression models control complexity by imposing a penalty on model parameters. By far, the most popular model in this group is *Lasso* (least absolute shrinkage and selection operator) proposed by Tibshirani (1996). In this approach, the coefficients are obtained by minimizing the residual sum of squares (RSS) subject to a constraint on the magnitude of regression coefficients,

$$\min_{\|\beta\|_1 \leq t} \text{RSS}(\beta), \qquad \text{RSS}(\beta) := \sum_i (y_i - \beta_0 - x_i^\mathsf{T}\beta)^2 \tag{45}$$

One could estimate the parameters by solving the following optimization problem:

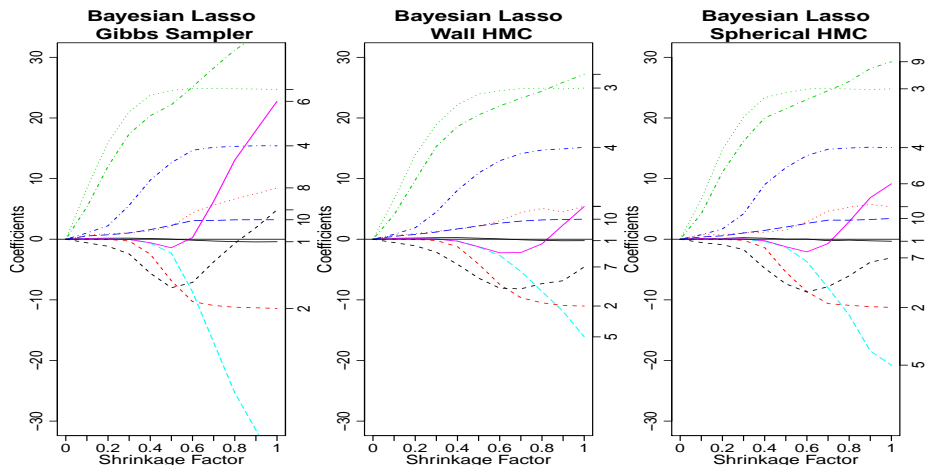$$\min_{\beta,\lambda} \text{RSS}(\beta) + \lambda\|\beta\|_1 \tag{46}$$

Figure 8: Bayesian Lasso using three different sampling algorithms: Gibbs sampler (left), Wall HMC (middle) and Spherical HMC (right).

where $\lambda \geq 0$ is the regularization parameter. Park and Casella (2008) and Hans (2009) have proposed a Bayesian alternative method, called Bayesian Lasso, where the penalty term is replaced by a prior distribution of the form $P(\beta) \propto \exp(-\lambda|\beta|)$, which can be represented as a scale mixture of normal distributions (West, 1987). This leads to a hierarchical Bayesian model with full conditional conjugacy; therefore, the Gibbs sampler can be used for inference.

Our proposed spherical augmentation in this paper can directly handle the constraints in Lasso models. That is, we can *conveniently* use Gaussian priors for model parameters, $\beta|\sigma^2 \sim \mathcal{N}(0, \sigma^2 I)$, and let the sampler *automatically* handle the constraint. In particular, c-SphHMC can be used to sample posterior distribution of β with the 1-norm constraint. For this problem, we modify the Wall HMC algorithm, which was originally proposed for box type constraints (Neal, 2011). See Appendix E for more details.

We evaluate our method based on the diabetes data set (N=442, D=10) discussed in Park and Casella (2008). Figure 8 compares coefficient estimates given by the Gibbs sampler (Park and Casella, 2008), Wall HMC, and Spherical HMC respectively as the shrinkage factor $s := \|\hat{\beta}^{\text{Lasso}}\|_1/\|\hat{\beta}^{\text{OLS}}\|_1$ changes from 0 to 1. Here, $\hat{\beta}^{\text{OLS}}$ denotes the estimates obtained by ordinary least squares (OLS) regression. For the Gibbs sampler, we choose different λ so that the corresponding shrinkage factor s varies from 0 to 1. For Wall HMC and Spherical HMC, we fix the number of leapfrog steps to 10 and set the trajectory length such that they both have comparable acceptance rates around 70%.

Figure 9 compares the sampling efficiency of these three methods. As we impose tighter constraints (i.e., lower shrinkage factors s), Spherical HMC becomes substantially more efficient than the Gibbs sampler and Wall HMC.
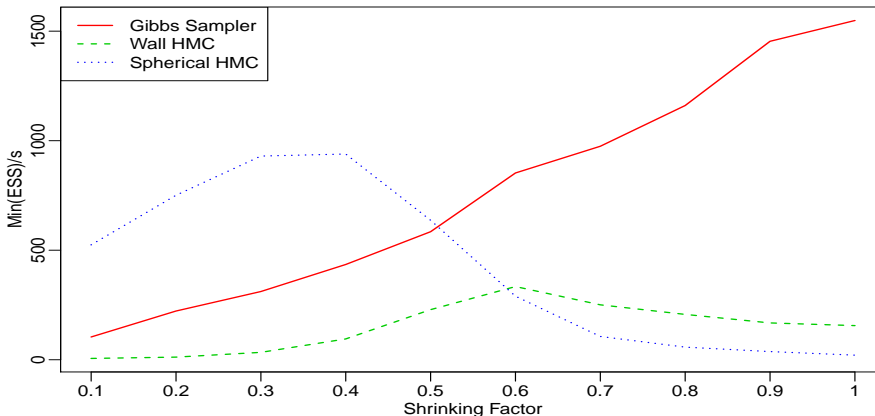
Figure 9: Sampling efficiency of different algorithms for Bayesian Lasso based on the diabetes dataset.

5.3 Bridge regression

The Lasso model discussed in the previous section is in fact a member of a family of regression models called *Bridge regression* (Frank and Friedman, 1993), where the coefficients are obtained by minimizing the residual sum of squares subject to a constraint on the magnitude of regression coefficients as follows:

$$\min_{\|\beta\|_q \leq t} \text{RSS}(\beta), \qquad \text{RSS}(\beta) := \sum_i (y_i - \beta_0 - x_i^\mathsf{T}\beta)^2 \tag{47}$$

For Lasso, $q = 1$, which allows the model to force some of the coefficients to become exactly zero (i.e., become excluded from the model). When $q = 2$, this model is known as *ridge regression*. Bridge regression is more flexible by allowing different q norm constraints for different effects on shrinking the magnitude of parameters (See Figure 10).

While the Gibbs sampler method of Park and Casella (2008) and Hans (2009) is limited to Lasso, our approach can be applied to all bridge regression models with different q. To handle the general q-norm constraint, one can map the constrained domain to the unit ball by (13) and apply c-SphHMC. Figure 10 compares the parameter estimates of Bayesian Lasso to the estimates obtained from two Bridge regression models with $q = 1.2$ and $q = 0.8$ for the diabetes dataset (Park and Casella, 2008) using our Spherical HMC algorithm. As expected, tighter constraints (e.g., $q = 0.8$) would lead to faster shrinkage of regression parameters as we decrease s.

5.4 Reconstruction of quantized stationary Gaussian process

We now investigate the example of reconstructing quantized stationary Gaussian process discussed in Pakman and Paninski (2013). Suppose we are given N values of a function $f(x_i), i = 1, \cdots, N$, which takes discrete values from $\{q_k\}_{k=1}^K$. We assume that this is a
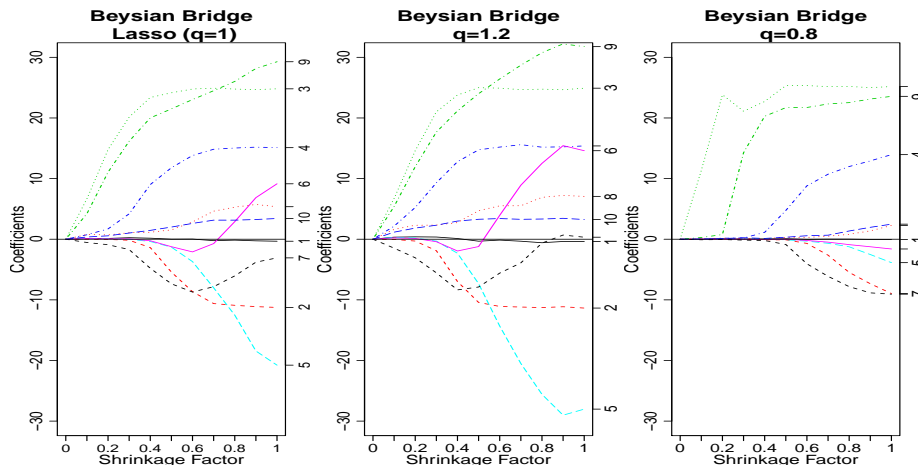
22

Figure 10: Bayesian Bridge Regression by Spherical HMC: Lasso (q=1, left), q=1.2 (middle), and q=0.8 (right).

quantized projection of a sample $y(x_i)$ from a stationary Gaussian process with a known translation-invariant covariance kernel of the form $\Sigma_{ij} = K(|x_i - x_j|)$, and the quantization follows a known rule of the form

$$f(x_i) = q_k, \quad \text{if } z_k \le y(x_i) < z_{k+1} \tag{48}$$

The objective is to sample from the posterior distribution

$$p(y(x_1), \cdots, y(x_N)|f(x_1), \cdots, f(x_N)) \sim \mathcal{N}(0, \Sigma) \quad \text{trunctated by rule (48)} \tag{49}$$

In this example, the function is sampled from a Gaussian process with the following kernel

$$K(|x_i - x_j|) = \sigma^2 \exp\left\{-\frac{|x_i - x_j|^2}{2\eta^2}\right\}, \quad \sigma^2 = 0.6, \ \eta^2 = 0.2$$

We sample $N = 100$ points of $\{y(x_i)\}$ and quantize them with

$q_1 = -0.75, \ q_2 = -0.25, \ q_3 = 0.25, \ q_4 = 0.75, \ z_1 = -\infty, \ z_2 = -0.5, \ z_3 = 0, \ z_4 = 0.5, \ z_5 = +\infty$

This example involves two types of constraints: box type (two sided) constraints and one sided constraints. In implementing our Spherical HMC algorithms, we transform the subspace formed by components with both finite lower and upper limits into unit ball and map the subspace formed by components with one sided constraints to the whole space using absolute value (discussed at the end of Section 3).

Figure 11 shows the quantized Gaussian process (upper) and the estimates (lower) with 10^5 samples given by different MCMC algorithms. Overall, all the methods recover the truth well. Table 3 summarizes the efficiency of sampling 1.1×10^5 and burning the first 10^4 with RWM, Wall HMC, exact HMC, c-SphHMC and s-SphHMC. Exact HMC generates more effective samples but takes much longer time even though implemented in C. Spherical HMC algorithms outperform it in terms of time normalized ESS. Interestingly, Wall HMC performs well in this example, even better than exact HMC and c-SphHMC.
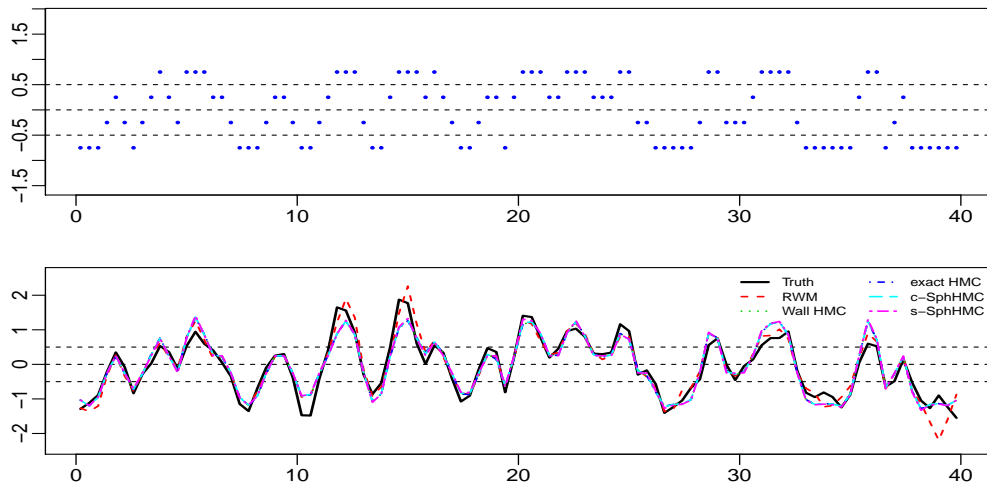
Figure 11: Quantized stationary Gaussian process (upper) and the estimates of the process (lower).

Method	AP	s/iter	ESS(min,med,max)	Min(ESS)/s	spdup
RWM	0.70	7.11E-05	(2,9,35)	0.22	1.00
Wall HMC	0.69	9.94E-04	(12564,24317,43876)	114.92	534.48
exact HMC	1.00	1.00E-02	(72074,1e+05,1e+05)	65.31	303.76
c-SphHMC	0.72	1.73E-03	(13029,26021,56445)	68.44	318.32
s-SphHMC	0.80	1.09E-03	(14422,31182,81948)	120.59	560.86

Table 3: Comparing efficiency of RWM, Wall HMC, exact HMC, c-SphHMC and s-SphHMC in reconstructing a quantized stationary Gaussian process. AP is acceptance probability, s/iter is seconds per iteration, ESS(min,med,max) is the (minimal,median,maximal) effective sample size, and Min(ESS)/s is the minimal ESS per second.

5.5 LDA on Wikipedia corpus

LDA (Blei et al., 2003) is a popular hierarchical Bayesian model for topic modeling. The model consists of K topics with probabilities $\{\pi_k\}$ drawn from a symmetric Dirichlet prior $\mathrm{Dir}(\beta)$. A document d is modeled by a mixture of topics, with mixing proportions $\eta_d \sim \mathrm{Dir}(\alpha)$. Document d is assumed to be generated by i.i.d. sampling of a topic assignment, z_{di}, from η_d for each word w_{di} in the document, and then drawing the word w_{di} from the assigned topic with probability $\pi_{z_{di}}$ (Patterson and Teh, 2013). Teh et al. (2006) integrate out η analytically to obtain the following semi-collapsed distribution:

$$p(w, z, \pi | \alpha, \beta) = \prod_{d=1}^{D} \frac{\Gamma(K\alpha)}{\Gamma(K\alpha + n_{d..})} \prod_{k=1}^{K} \frac{\Gamma(\alpha + n_{dk.})}{\Gamma(\alpha)} \prod_{k=1}^{K} \frac{\Gamma(W\beta)}{\Gamma(\beta)^W} \prod_{w=1}^{W} \pi_{kw}^{\beta + n_{.kw} - 1} \tag{50}$$
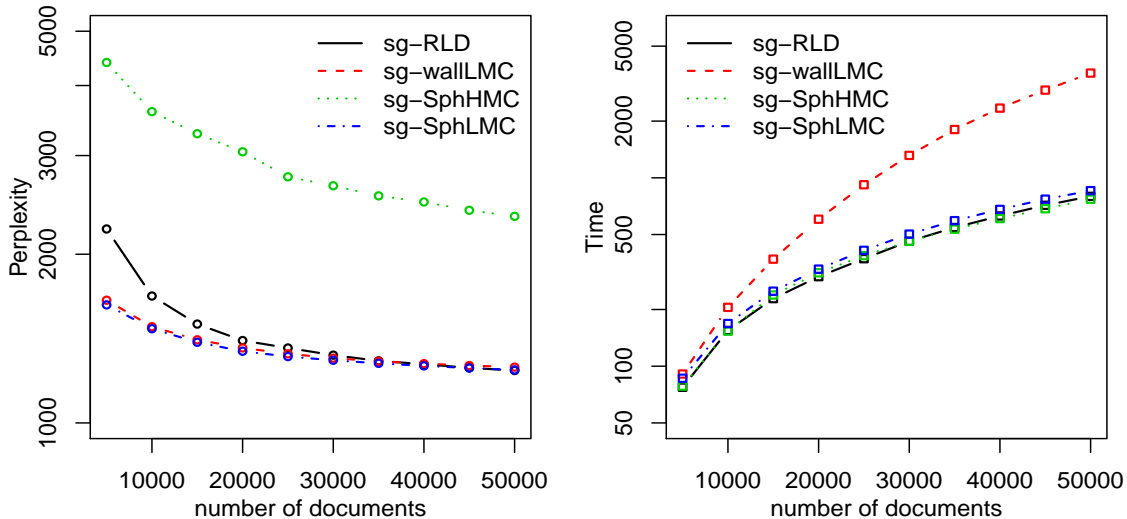
Figure 12: Test-set perplexity and computation time (in log scale) based on the Wikipedia corpus.

where $n_{dkw} = \sum_{i=1}^{N_d} \delta(w_{di} = w, z_{di} = k)$. Here, "·" denotes the summation over the corresponding index. Given π, the documents are i.i.d so the above equation can be factorized as follows (Patterson and Teh, 2013):

$$p(w, z, \pi | \alpha, \beta) = p(\pi | \beta) \prod_{d=1}^{D} p(w_d, z_d | \alpha, \pi), \quad p(w_d, z_d | \alpha, \pi) = \prod_{k=1}^{K} \frac{\Gamma(\alpha + n_{dk\cdot})}{\Gamma(\alpha)} \prod_{w=1}^{W} \pi_{kw}^{n_{dkw}}$$
$$(51)$$

To evaluate our proposed methods, we compare them with the state-of-the-art method of Patterson and Teh (2013). Their approach, called stochastic gradient Riemannian Langevin dynamics (sg-RLD) is an extension of the stochastic gradient Langevin dynamics (SGLD) proposed by Welling and Teh (2011). Because this approach uses mini-batches of data to approximate the gradient and omits the accept/reject step of Metropolis-Hastings while decreasing the step size, we follow the same procedure to make our methods comparable. Further, because Langevin dynamics can be regarded as a single step Hamiltonian dynamics (Neal, 2011), we set $L = 1$. We refer the resulting algorithms as sg-SphHMC and sg-SphLMC, which are modified versions of our SphHMC and SphLMC algorithms. sg-SphLMC uses the following stochastic (natural) gradient (gradient preconditioned with

25

metric)

$$g_{kw} = [(n_{kw}^* + \beta - 1/2)/\theta_{kw} + \theta_{kw}(n_{k\cdot}^* + W(\beta - 1/2))]/(2 * n_{k\cdot}^*), \quad n_{kw}^* = \frac{|D|}{|D_t|} \sum_{d \in D_t} \mathrm{E}_{z_d|w_d, \theta, \alpha}[n_{dkw}]$$

$$(52)$$

where $1/2$ comes from the logarithm of volume adjustment. In contrast, the stochastic gradient for sg-SphHMC is $4g_{kw}n_{k\cdot}^*$. (See Section 4.3). The expectation in Equation (52) is calculated using Gibbs sampling on the topic assignment in each document separately, given the conditional distributions (Patterson and Teh, 2013)

$$p(z_{di} = k|w_d, \theta, \alpha) = \frac{(\alpha + n_{dk\cdot}^{\backslash i})\pi_{kw_{di}}}{\sum_k (\alpha + n_{dk\cdot}^{\backslash i})\pi_{kw_{di}}} \tag{53}$$

where $\backslash i$ means a count excluding the topic assignment variable currently being updated. Step size is decreased according to $\varepsilon_t = a(1 + t/b)^{-c}$.

We use perplexity (Patterson and Teh, 2013; Wallach et al., 2009) to compare the predictive performance of different methods in terms of the probability they assign to unseen data,

$$\mathrm{perp}(w_d|\mathcal{W}, \alpha, \beta) = \exp\left\{-\sum_{i=1}^{n_{d\cdot\cdot}} \log p(w_{di}|\mathcal{W}, \alpha, \beta)/n_{d\cdot\cdot}\right\}, \quad p(w_{di}|\mathcal{W}, \alpha, \beta) = \mathrm{E}_{\eta_d, \pi}[\sum_k \eta_{dk}\pi_{kw_{di}}]$$

$$(54)$$

where $\mathcal{W}$ is the training set and w_d is the hold-out sample. More specifically, we use the document completion approach (Wallach et al., 2009), which partitions the test document w_d into two sets, w_d^{train} and w_d^{test}; we then use w_d^{train} to estimate n_d for the test document and use w_d^{test} to calculate perplexity.

We train the model online using 50000 documents randomly downloaded from Wikipedia with the vocabulary of approximately 8000 words created from Project Gutenburg texts (Hoffman et al., 2010). The perplexity is evaluated on 1000 held-out documents. A mini-batch of 50 documents is used for updating the natural gradient for 4 algorithms: sg-RLD, sg-wallLMC[3], sg-SphHMCand sg-SphLMC.

Figure 12 compares the above methods in terms of their perplexities. For each method, we show the best performance over different settings (Settings for best performance are listed in Table 4.). Both sg-wallLMC and sg-SphLMC have lower perplexity than sg-RLD at early stage, when relatively a small number of documents are used for training; as the number of training documents increases, the methods reach the same level of performance. As expected, sg-SphHMC does not perform well due to the absence of a proper scaling provided by the Fisher metric.

6. Discussion

We have introduced a new approach, *spherical augmentation*, for sampling from constrained probability distributions. This method maps the constrained domain to a sphere in an augmented space. Sampling algorithms can freely explore the surface of sphere to generate

---

3. The stochastic gradient for sg-wallLMC is $[(n_{kw}^* + \beta - 1/2) + \pi_{kw}(n_{k\cdot}^* + W(\beta - 1/2))]/n_{k\cdot}^*$.

Algorithm	a	b	c	α	β	K	Gibbs samples
sg-RLD	0.01	1000	0.6	0.01	0.5000	100	100
sg-wallLMC	0.20	1000	2.0	0.01	0.5000	100	100
sg-SphHMC	0.01	1000	0.6	0.01	0.0100	100	100
sg-SphLMC	0.25	1000	1.5	0.01	0.5000	100	100

Table 4: Parameter settings for best performance in Wikipedia experiment.

samples that remain within the constrained domain when mapped back to the original space. This way, our proposed method provides a mathematically natural and computationally efficient framework that can be applied to a wide range of statistical inference problems with norm constraints.

The augmentation approach proposed here is based on the change of variables theorem. We augment the original D-dimensional space with one extra dimension by either inserting slack variables (c-SphHMC) or using embedding map (s-SphHMC), The augmented Hamiltonian is the same under different representations (30)(40) due to the mathematical fact that the energy is invariant to the choice of coordinates (Proposition A.1). To account for the change of geometry, a volume adjustment term needs to be used, either as a weight after obtaining all the samples (SphHMC) or as an added term to the total energy (SphLMC).

Our proposed method takes advantage of the splitting strategy to further improve computational efficiency. We split the Lagrangian dynamics and update velocity in the tangent space, rather than momentum in the cotangent space. This implementation avoids the requirement of embedding as in Byrne and Girolami (2013) and could be applied to more general situations.

In developing Spherical HMC, we start with the standard HMC, using the Euclidean metric $\mathbf{I}$ on unit ball $\mathbf{B}_0^D(1)$. Then, spherical geometry is introduced to handle constraints. One possible future direction could be to directly start with RHMC/LMC, which use a more informative metric (i.e., the Fisher metric $\mathbf{G_F}$), and then incorporate the spherical geometry for the constraints. For example, a possible metric for the augmented space could be $\mathbf{G_F} + \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}/\theta_{D+1}^2$. However, under such a metric, we might not be able to find the geodesic flow analytically, which could undermine the added benefit from using the Fisher metric.

In future, we also intend to explore the possibility of applying the spherical augmentation to Elliptical Slice sampler (Murray et al., 2010) in order to generalize it to Spherical Slice sampler (SSS). The resulting algorithm can be applied to truncated Gaussian process models. In general, we can extend our proposed methods to infinite dimensional function spaces. This would involve the infinite dimensional manifold $\mathcal{S}^\infty := \{f \in L^2(\Omega)| \int f^2 d\mu = 1\}$. In this setting it is crucial to ensure that the acceptance probability does not drop quickly as dimension increases (Beskos et al., 2011).

Appendix A. Spherical Geometry

We first discuss the geometry of the D-dimensional sphere $\mathcal{S}^D := \{\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{D+1} : \|\tilde{\boldsymbol{\theta}}\|_2 = 1\} \hookrightarrow \mathbb{R}^{D+1}$ under different coordinate systems, namely, the *Cartesian coordinate* and the *spherical coordinate*. Since $\mathcal{S}^D$ can be embedded (injectively and differentiably mapped to) in $\mathbb{R}^{D+1}$, we first introduce the concept of 'induced metric'.

Definition 1 (induced metric) *If $\mathcal{D}^d$ can be embedded to $\mathcal{M}^m$ $(m > d)$ by $f : U \subset \mathcal{D} \hookrightarrow \mathcal{M}$, then one can define the* induced metric, $g_{\mathcal{D}}$, *on $T\mathcal{D}$ through the metric $g_{\mathcal{M}}$ defined on $T\mathcal{M}$:*

$$g_{\mathcal{D}}(\boldsymbol{\theta})(\mathbf{u}, \mathbf{v}) = g_{\mathcal{M}}(f(\boldsymbol{\theta}))(df_{\boldsymbol{\theta}}(\mathbf{u}), df_{\boldsymbol{\theta}}(\mathbf{v})), \quad \mathbf{u}, \mathbf{v} \in T_{\boldsymbol{\theta}}\mathcal{D} \tag{55}$$

Remark 1 *For any $f : U \subset \mathcal{S}^D \hookrightarrow \mathbb{R}^{D+1}$, we can define the induced metric through dot product on $\mathbb{R}^{D+1}$. More specifically,*

$$g_{\mathcal{S}}(\mathbf{u}, \mathbf{v}) = [(Df)\mathbf{u}]^{\mathsf{T}}(Df)\mathbf{v} = \mathbf{u}^{\mathsf{T}}[(Df)^{\mathsf{T}}(Df)]\mathbf{v} \tag{56}$$

where $(Df)_{(D+1)\times D}$ is the Jacobian matrix of the mapping f. A Metric induced from dot product on Euclidean space is called a "canonical metric". This observation leads to the following simple fact that lays down the foundation of Spherical HMC.

Proposition A.1 (Energy invariance) *Kinetic energy $\frac{1}{2}\langle \mathbf{v}, \mathbf{v} \rangle_{\mathbf{G}(\boldsymbol{\theta})}$ is invariant to the choice of coordinate systems.*

Proof For any $\mathbf{v} \in T_{\boldsymbol{\theta}}\mathcal{D}$, suppose $\boldsymbol{\theta}(t)$ such that $\dot{\boldsymbol{\theta}}(0) = \mathbf{v}$. Denote the pushforward of $\mathbf{v}$ by embedding map $f : \mathcal{D} \to \mathcal{M}$ as $\tilde{\mathbf{v}} := f_*(\mathbf{v}) = \frac{d}{dt}(f \circ \boldsymbol{\theta})(0)$. Then we have

$$\frac{1}{2}\langle \mathbf{v}, \mathbf{v} \rangle_{\mathbf{G}(\boldsymbol{\theta})} = \frac{1}{2}g_{\mathcal{M}}(f(\boldsymbol{\theta}))(\tilde{\mathbf{v}}, \tilde{\mathbf{v}}) \tag{57}$$

That is, regardless of the form of the energy under a coordinate system, its value is the same as the one in the embedded manifold. In particular, when $\mathcal{M} = \mathbb{R}^{D+1}$, the right hand side simplifies to $\frac{1}{2}\|\tilde{\mathbf{v}}\|_2^2$. $\blacksquare$

A.1 Canonical metric in the Cartesian coordinate

Now consider the D-dimensional ball $\mathcal{B}_{\mathbf{0}}^D(1) := \{\boldsymbol{\theta} \in \mathbb{R}^D : \|\boldsymbol{\theta}\|_2 \leq 1\}$. Here, $\{\boldsymbol{\theta}, \mathcal{B}_{\mathbf{0}}^D(1)\}$ can be viewed as the Cartesian coordinate system for $\mathcal{S}^D$. The coordinate mapping $T_{\mathcal{B} \to \mathcal{S}_+} : \boldsymbol{\theta} \mapsto \tilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}, \theta_{D+1})$ in (6) can be viewed as the embedding map into $\mathbb{R}^{D+1}$, and the Jacobian matrix of $T_{\mathcal{B} \to \mathcal{S}_+}$ is $dT_{\mathcal{B} \to \mathcal{S}_+} = \frac{d\tilde{\boldsymbol{\theta}}}{d\boldsymbol{\theta}^{\mathsf{T}}} = \begin{bmatrix} \mathbf{I}_D \\ -\boldsymbol{\theta}^{\mathsf{T}}/\theta_{D+1} \end{bmatrix}$. Therefore the *canonical metric* of $\mathcal{S}^D$ in the Cartesian coordinate, $\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})$, is

$$\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta}) = dT_{\mathcal{B} \to \mathcal{S}_+}^{\mathsf{T}} dT_{\mathcal{B} \to \mathcal{S}_+} = \mathbf{I}_D + \frac{\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}}{\theta_{D+1}^2} = \mathbf{I}_D + \frac{\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}}{1 - \|\boldsymbol{\theta}\|_2^2} \tag{58}$$

Another way to obtain the metric is through the first fundamental form ds^2 (i.e., squared infinitesimal length of a curve) for $\mathcal{S}^D$, which can be expressed in terms of the differential form $d\boldsymbol{\theta}$ and the canonical metric $\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})$,

$$ds^2 = \langle d\boldsymbol{\theta}, d\boldsymbol{\theta} \rangle_{\mathbf{G}_{\mathcal{S}_c}} = d\boldsymbol{\theta}^\mathsf{T} \mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

On the other hand, ds^2 can also be obtained as follows (Spivak, 1979):

$$ds^2 = \sum_{i=1}^{D+1} d\theta_i^2 = \sum_{i=1}^{D} d\theta_i^2 + (d(\theta_{D+1}(\boldsymbol{\theta})))^2 = d\boldsymbol{\theta}^\mathsf{T} d\boldsymbol{\theta} + \frac{(\boldsymbol{\theta}^\mathsf{T} d\boldsymbol{\theta})^2}{1 - \|\boldsymbol{\theta}\|_2^2} = d\boldsymbol{\theta}^\mathsf{T}[\mathbf{I} + \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}/\theta_{D+1}^2] d\boldsymbol{\theta}$$

Equating the above two quantities yields the form of the canonical metric $\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})$ as in Equation (58). This viewpoint provides a natural way to explain the length of tangent vector. For any vector $\tilde{\mathbf{v}} = (\mathbf{v}, v_{D+1}) \in T_{\tilde{\boldsymbol{\theta}}}\mathcal{S}^D = \{\tilde{\mathbf{v}} \in \mathbb{R}^{D+1} : \tilde{\boldsymbol{\theta}}^\mathsf{T}\tilde{\mathbf{v}} = 0\}$, one could think of $\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})$ as a mean to express the length of $\tilde{\mathbf{v}}$ in terms of $\mathbf{v}$,

$$\mathbf{v}^\mathsf{T} \mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta}) \mathbf{v} = \|\mathbf{v}\|_2^2 + \frac{\mathbf{v}^\mathsf{T} \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T} \mathbf{v}}{\theta_{D+1}^2} = \|\mathbf{v}\|_2^2 + \frac{(-\theta_{D+1}v_{D+1})^2}{\theta_{D+1}^2} = \|\mathbf{v}\|_2^2 + v_{D+1}^2 = \|\tilde{\mathbf{v}}\|_2^2 \quad (59)$$

This indeed verifies the energy invariance Proposition A.1.

The following proposition provides the analytic forms of the determinant and the inverse of $\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})$.

Proposition A.2 *The determinant and the inverse of the* canonical metric *are as follows*

$$|\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})| = \theta_{D+1}^{-2}, \qquad \mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})^{-1} = \mathbf{I}_D - \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T} \quad (60)$$

Proof The determinant of the canonical metric $\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})$ is given by the matrix determinant lemma,

$$|\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})| = \det\left[\mathbf{I}_D + \frac{\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}}{\theta_{D+1}^2}\right] = 1 + \frac{\boldsymbol{\theta}^\mathsf{T}\boldsymbol{\theta}}{\theta_{D+1}^2} = \frac{1}{\theta_{D+1}^2}$$

The inverse of $\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})$ is obtained by the Sherman-Morrison-Woodbury formula (Golub and Van Loan, 1996)

$$\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})^{-1} = \left[\mathbf{I}_D + \frac{\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}}{\theta_{D+1}^2}\right]^{-1} = \mathbf{I}_D - \frac{\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}/\theta_{D+1}^2}{1 + \boldsymbol{\theta}^\mathsf{T}\boldsymbol{\theta}/\theta_{D+1}^2} = \mathbf{I}_D - \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}$$

∎

Corollary 1 *The volume adjustment of changing measure in* (7) *is*

$$\left|\frac{d\boldsymbol{\theta}_\mathcal{B}}{d\boldsymbol{\theta}_{\mathcal{S}_c}}\right| = |\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})|^{-\frac{1}{2}} = |\theta_{D+1}| \quad (61)$$

Proof Canonical measure can be defined through the Riesz representation theorem by using a positive linear functional on the space $C_0(\mathcal{S}^D)$ of compactly supported continuous functions on $\mathcal{S}^D$ (Spivak, 1979; do Carmo, 1992). More precisely, there is a unique positive Borel measure μ_c such that for (any) coordinate chart $(\mathcal{B}_{\mathbf{0}}^D(1), T_{\mathcal{B} \to \mathcal{S}_+})$,

$$\int_{\mathcal{S}_+^D} f(\tilde{\boldsymbol{\theta}}) d\boldsymbol{\theta}_{\mathcal{S}_c} = \int_{\mathcal{B}_{\mathbf{0}}^D(1)} f(\boldsymbol{\theta}) \sqrt{|\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})|} d\boldsymbol{\theta}_{\mathcal{B}}$$

where $\mu_c = d\boldsymbol{\theta}_{\mathcal{S}_c}$, and $d\boldsymbol{\theta}_{\mathcal{B}}$ is the Euclidean measure. Therefore we have

$$\left| \frac{d\boldsymbol{\theta}_{\mathcal{S}_c}}{d\boldsymbol{\theta}_{\mathcal{B}}} \right| = |\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})|^{\frac{1}{2}} = |\theta_{D+1}|^{-1}$$

Alternatively, $\left| \frac{d\boldsymbol{\theta}_{\mathcal{B}}}{d\boldsymbol{\theta}_{\mathcal{S}_c}} \right| = |\theta_{D+1}|$. ∎

A.2 Geodesic on a sphere in the Cartesian coordinate

To find the geodesic on a sphere, we need to solve the following equations:

$$\dot{\boldsymbol{\theta}} = \mathbf{v} \tag{62}$$

$$\dot{\mathbf{v}} = -\mathbf{v}^{\mathsf{T}} \boldsymbol{\Gamma}_{\mathcal{S}_c}(\boldsymbol{\theta}) \mathbf{v} \tag{63}$$

for which we need to calculate the Christoffel symbols, $\boldsymbol{\Gamma}_{\mathcal{S}_c}(\boldsymbol{\theta})$, first. Note that the (i,j)-th element of $\mathbf{G}_{\mathcal{S}_c}$ is $g_{ij} = \delta_{ij} + \theta_i \theta_j / \theta_{D+1}^2$, and the (i,j,k)-th element of $d\mathbf{G}_{\mathcal{S}_c}$ is $g_{ij,k} = (\delta_{ik}\theta_j + \theta_i \delta_{jk})/\theta_{D+1}^2 + 2\theta_i\theta_j\theta_k/\theta_{D+1}^4$. Therefore

$$\begin{aligned} \Gamma_{ij}^k &= \frac{1}{2} g^{kl}[g_{lj,i} + g_{il,j} - g_{ij,l}] \\ &= \frac{1}{2}(\delta^{kl} - \theta^k\theta^l)[(\delta_{li}\theta_j + \theta_l\delta_{ji})/\theta_{D+1}^2 + (\delta_{ij}\theta_l + \theta_i\delta_{lj})/\theta_{D+1}^2 - (\delta_{il}\theta_j + \theta_i\delta_{jl})/\theta_{D+1}^2 + 2\theta_i\theta_j\theta_l/\theta_{D+1}^4] \\ &= (\delta^{kl} - \theta^k\theta^l)\theta_l/\theta_{D+1}^2[\delta_{ij} + \theta_i\theta_j/\theta_{D+1}^2] \\ &= \theta_k[\delta_{ij} + \theta_i\theta_j/\theta_{D+1}^2] = [\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta}) \otimes \boldsymbol{\theta}]_{ijk} \end{aligned}$$

Using these results, we can write Equation (63) as $\dot{\mathbf{v}} = -\mathbf{v}^{\mathsf{T}} \mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta}) \mathbf{v} \boldsymbol{\theta} = -\|\tilde{\mathbf{v}}\|_2^2 \boldsymbol{\theta}$. Further, we have

$$\dot{\theta}_{D+1} = \frac{d}{dt}\sqrt{1 - \|\boldsymbol{\theta}\|_2^2} = -\frac{\boldsymbol{\theta}^{\mathsf{T}}}{\theta_{D+1}}\dot{\boldsymbol{\theta}} \qquad\qquad = v_{D+1}$$

$$\dot{v}_{D+1} = -\frac{d}{dt}\frac{\boldsymbol{\theta}^{\mathsf{T}}\mathbf{v}}{\theta_{D+1}} \quad = -\frac{\dot{\boldsymbol{\theta}}^{\mathsf{T}}\mathbf{v} + \boldsymbol{\theta}^{\mathsf{T}}\dot{\mathbf{v}}}{\theta_{D+1}} + \frac{\boldsymbol{\theta}^{\mathsf{T}}\mathbf{v}}{\theta_{D+1}^2}\dot{\theta}_{D+1} = -\|\tilde{\mathbf{v}}\|_2^2\theta_{D+1}$$

Therefore, we can rewrite the geodesic equations (62)(63) with augmented components as

$$\dot{\tilde{\boldsymbol{\theta}}} = \tilde{\mathbf{v}} \tag{64}$$

$$\dot{\tilde{\mathbf{v}}} = -\|\tilde{\mathbf{v}}\|_2^2 \tilde{\boldsymbol{\theta}} \tag{65}$$

Multiplying both sides of Equation (65) by $\tilde{\mathbf{v}}^{\mathsf{T}}$ to obtain $\frac{d}{dt}\|\tilde{\mathbf{v}}\|_2^2 = 0$, we can solve the above system of differential equations as follows:

$$\tilde{\boldsymbol{\theta}}(t) = \tilde{\boldsymbol{\theta}}(0)\cos(\|\tilde{\mathbf{v}}(0)\|_2 t) + \frac{\tilde{\mathbf{v}}(0)}{\|\tilde{\mathbf{v}}(0)\|_2}\sin(\|\tilde{\mathbf{v}}(0)\|_2 t)$$

$$\tilde{\mathbf{v}}(t) = -\tilde{\boldsymbol{\theta}}(0)\|\tilde{\mathbf{v}}(0)\|_2\sin(\|\tilde{\mathbf{v}}(0)\|_2 t) + \tilde{\mathbf{v}}(0)\cos(\|\tilde{\mathbf{v}}(0)\|_2 t)$$

A.3 Round metric in the spherical coordinate

Consider the D-dimensional hyper-rectangle $\mathcal{R}_0^D := [0,\pi]^{D-1} \times [0,2\pi)$ and the corresponding spherical coordinate system, $\{\boldsymbol{\theta}, \mathcal{R}_0^D\}$, for $\mathcal{S}^D$. The coordinate mapping $T_{\mathcal{R}_0 \to \mathcal{S}} : \boldsymbol{\theta} \mapsto \mathbf{x}$, $x_d = \cos(\theta_d)\prod_{i=1}^{d-1}\sin(\theta_i)$, $d = 1, \cdots, D+1$, $(\theta_{D+1} = 0)$ can be viewed as the embedding map into $\mathbb{R}^{D+1}$, and the Jacobian matrix of $T_{\mathcal{R}_0 \to \mathcal{S}}$ is $\frac{d\mathbf{x}}{d\boldsymbol{\theta}^{\mathsf{T}}}$ with the (d,j)-th element $[-\tan(\theta_d)\delta_{dj} + \cot(\theta_j)I(j < d)]x_d$. The induced metric of $\mathcal{S}^D$ in the spherical coordinate is called *round metric*, denoted as $\mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta})$, whose (i,j)-th element is as follows

$$\mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta})_{ij}$$
$$= \sum_{d=1}^{D+1}[-\tan(\theta_d)\delta_{di} + \cot(\theta_j)I(i < d)][-\tan(\theta_d)\delta_{dj} + \cot(\theta_j)I(j < d)]x_d^2$$
$$= \tan^2(\theta_i)\delta_{ij}x_i^2 - \tan(\theta_i)\cot(\theta_j)I(i > j)x_i^2 - \tan(\theta_j)\cot(\theta_i)I(i < j)x_j^2 + \cot(\theta_i)\cot(\theta_j)\sum_{d>\max\{i,j\}}x_d^2$$
$$= \begin{cases} -\tan(\theta_j)\cot(\theta_i)x_j^2 + \cot(\theta_i)\cot(\theta_j)\sum_{d>j}x_d^2 = 0, & i < j \\ \tan^2(\theta_i)x_i^2 + \cot^2(\theta_i)\sum_{d>i}x_d^2 = (\tan^2(\theta_i) + 1)x_i^2 = \prod_{i=1}^{d-1}\sin^2(\theta_i), & i = j \end{cases}$$
$$= \prod_{i=1}^{d-1}\sin^2(\theta_i)\delta_{ij}$$

$$(66)$$

Therefore, $\mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta}) = \text{diag}[1, \sin^2(\theta_1), \cdots, \prod_{d=1}^{D-1}\sin^2(\theta_d)]$. Another way to obtain $\mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta})$ is through the coordinate change:

$$\mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta}) = \frac{d\boldsymbol{\theta}_{\mathcal{S}_c}^{\mathsf{T}}}{d\boldsymbol{\theta}_{\mathcal{S}_r}}\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})\frac{d\boldsymbol{\theta}_{\mathcal{S}_c}}{d\boldsymbol{\theta}_{\mathcal{S}_r}^{\mathsf{T}}} \tag{67}$$

Similar to Corollary (1), we have

Proposition A.3 *The volume adjustment of changing measure in* (10) *is*

$$\left|\frac{d\boldsymbol{\theta}_{\mathcal{R}_0}}{d\boldsymbol{\theta}_{\mathcal{S}_r}}\right| = |\mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta})|^{-\frac{1}{2}} = \prod_{d=1}^{D-1}\sin^{-(D-d)}(\theta_d) \tag{68}$$

Appendix B. Jacobian of the transformation between q-norm domains

The following proposition gives the weights needed for the transformation from $\mathcal{Q}^D$ to $\mathcal{B}_0^D(1)$.

Proposition B.1 *The Jacobian determinant (weight) of $T_{\mathcal{B} \to \mathcal{Q}}$ is as follows:*

$$|dT_{\mathcal{S} \to \mathcal{Q}}| = \left(\frac{2}{q}\right)^D \left(\prod_{i=1}^{D} |\theta_i|\right)^{2/q-1} \tag{69}$$

Proof Note

$$T_{\mathcal{B} \to \mathcal{Q}} : \boldsymbol{\theta} \mapsto \boldsymbol{\beta} = \mathrm{sgn}(\boldsymbol{\theta})|\boldsymbol{\theta}|^{2/q}$$

The Jacobian matrix for $T_{\mathcal{B} \to \mathcal{Q}}$ is

$$\frac{d\boldsymbol{\beta}}{d\boldsymbol{\theta}^{\mathsf{T}}} = \frac{2}{q}\mathrm{diag}(|\boldsymbol{\theta}|^{2/q-1})$$

Therefore the Jacobian determinant of $T_{\mathcal{B} \to \mathcal{Q}}$ is

$$|dT_{\mathcal{B} \to \mathcal{Q}}| = \left|\frac{d\boldsymbol{\beta}}{d\boldsymbol{\theta}^{\mathsf{T}}}\right| = \left(\frac{2}{q}\right)^D \left(\prod_{i=1}^{D} |\theta_i|\right)^{2/q-1}$$

∎

The following proposition gives the weights needed for the change of domains from $\mathcal{R}^D$ to $\mathcal{B}_{\mathbf{0}}^D(1)$.

Proposition B.2 *The Jacobian determinant (weight) of $T_{\mathcal{B} \to \mathcal{R}}$ is as follows:*

$$|dT_{\mathcal{B} \to \mathcal{R}}| = \frac{\|\boldsymbol{\theta}\|_2^D}{\|\boldsymbol{\theta}\|_\infty^D} \prod_{i=1}^{D} \frac{u_i - l_i}{2} \tag{70}$$

Proof First, we note

$$T_{\mathcal{B} \to \mathcal{R}} = T_{\mathcal{C} \to \mathcal{R}} \circ T_{\mathcal{B} \to \mathcal{C}} : \boldsymbol{\theta} \mapsto \boldsymbol{\beta}' = \boldsymbol{\theta}\frac{\|\boldsymbol{\theta}\|_2}{\|\boldsymbol{\theta}\|_\infty} \mapsto \boldsymbol{\beta} = \frac{\mathbf{u} - \mathbf{l}}{2}\boldsymbol{\beta}' + \frac{\mathbf{u} + \mathbf{l}}{2}$$

The corresponding Jacobian matrices are

$$
\begin{aligned}
T_{\mathcal{B} \to \mathcal{C}} : \frac{d\boldsymbol{\beta}'}{d\boldsymbol{\theta}^{\mathsf{T}}} &= \frac{\|\boldsymbol{\theta}\|_2}{\|\boldsymbol{\theta}\|_\infty}\left[\mathbf{I} + \boldsymbol{\theta}\left(\frac{\boldsymbol{\theta}^{\mathsf{T}}}{\|\boldsymbol{\theta}\|_2^2} - \frac{\mathbf{e}_{\arg\max|\boldsymbol{\theta}|}^{\mathsf{T}}}{\boldsymbol{\theta}_{\arg\max|\boldsymbol{\theta}|}}\right)\right] \\
T_{\mathcal{C} \to \mathcal{R}} : \frac{d\boldsymbol{\beta}}{d(\boldsymbol{\beta}')^{\mathsf{T}}} &= \mathrm{diag}\left(\frac{\mathbf{u} - \mathbf{l}}{2}\right)
\end{aligned}
$$

where $\mathbf{e}_{\arg\max|\boldsymbol{\theta}|}$ is a vector with $(\arg\max|\boldsymbol{\theta}|)$-th element 1 and all others 0. Therefore,

$$|dT_{\mathcal{B} \to \mathcal{R}}| = |dT_{\mathcal{C} \to \mathcal{R}}|\,|dT_{\mathcal{B} \to \mathcal{C}}| = \left|\frac{d\boldsymbol{\beta}}{d(\boldsymbol{\beta}')^{\mathsf{T}}}\right|\left|\frac{d\boldsymbol{\beta}'}{d\boldsymbol{\theta}^{\mathsf{T}}}\right| = \frac{\|\boldsymbol{\theta}\|_2^D}{\|\boldsymbol{\theta}\|_\infty^D} \prod_{i=1}^{D} \frac{u_i - l_i}{2}$$

∎

Appendix C. Splitting Hamiltonian (Lagrangian) dynamics on $\mathcal{S}^D$

Splitting the Hamiltonian dynamics and its usefulness in improving HMC is a well-studied topic of research (Leimkuhler and Reich, 2004; Shahbaba et al., 2014; Byrne and Girolami, 2013). Splitting the Lagrangian dynamics (used in our approach), on the other hand, has not been discussed in the literature, to the best of our knowledge. Therefore, we prove the validity of our splitting method by starting with the well-understood method of splitting Hamiltonian (Byrne and Girolami, 2013),

$$H^*(\boldsymbol{\theta}, \mathbf{p}) = \frac{1}{2}U(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{p}^\mathsf{T}\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})^{-1}\mathbf{p} + \frac{1}{2}U(\boldsymbol{\theta})$$

The corresponding systems of differential equations,

$$\begin{cases} \dot{\boldsymbol{\theta}} &= \mathbf{0} \\ \dot{\mathbf{p}} &= -\frac{1}{2}\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta}) \end{cases} \qquad \begin{cases} \dot{\boldsymbol{\theta}} &= \mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})^{-1}\mathbf{p} \\ \dot{\mathbf{p}} &= -\frac{1}{2}\mathbf{p}^\mathsf{T}\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})^{-1}d\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})^{-1}\mathbf{p} \end{cases}$$

can be written in terms of Lagrangian dynamics in $(\boldsymbol{\theta}, \mathbf{v})$ as follows:

$$\begin{cases} \dot{\boldsymbol{\theta}} &= \mathbf{0} \\ \dot{\mathbf{v}} &= -\frac{1}{2}\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta}) \end{cases} \qquad \begin{cases} \dot{\boldsymbol{\theta}} &= \mathbf{v} \\ \dot{\mathbf{v}} &= -\mathbf{v}^\mathsf{T}\boldsymbol{\Gamma}_{\mathcal{S}_c}(\boldsymbol{\theta})\mathbf{v} \end{cases}$$

We have solved the second dynamics (on the right) in Section A.2. To solve the first dynamics, we note that

$$\dot{\theta}_{D+1} = \frac{d}{dt}\sqrt{1 - \|\boldsymbol{\theta}\|_2^2} = -\frac{\boldsymbol{\theta}^\mathsf{T}}{\theta_{D+1}}\dot{\boldsymbol{\theta}} \qquad\qquad = 0$$

$$\dot{v}_{D+1} = -\frac{d}{dt}\frac{\boldsymbol{\theta}^\mathsf{T}\mathbf{v}}{\theta_{D+1}} \quad = -\frac{\dot{\boldsymbol{\theta}}^\mathsf{T}\mathbf{v} + \boldsymbol{\theta}^\mathsf{T}\dot{\mathbf{v}}}{\theta_{D+1}} + \frac{\boldsymbol{\theta}^\mathsf{T}\mathbf{v}}{\theta_{D+1}^2}\dot{\theta}_{D+1} = \frac{1}{2}\frac{\boldsymbol{\theta}^\mathsf{T}}{\theta_{D+1}}\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta})$$

Therefore, we have

$$\tilde{\boldsymbol{\theta}}(t) = \tilde{\boldsymbol{\theta}}(0)$$

$$\tilde{\mathbf{v}}(t) = \tilde{\mathbf{v}}(0) - \frac{t}{2}\begin{bmatrix} \mathbf{I} \\ -\frac{\boldsymbol{\theta}(0)^\mathsf{T}}{\theta_{D+1}(0)} \end{bmatrix}[\mathbf{I} - \boldsymbol{\theta}(0)\boldsymbol{\theta}(0)^\mathsf{T}]\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta})$$

where $\begin{bmatrix} \mathbf{I} \\ -\frac{\boldsymbol{\theta}(0)^\mathsf{T}}{\theta_{D+1}(0)} \end{bmatrix}[\mathbf{I} - \boldsymbol{\theta}(0)\boldsymbol{\theta}(0)^\mathsf{T}] = \begin{bmatrix} \mathbf{I} - \boldsymbol{\theta}(0)\boldsymbol{\theta}(0)^\mathsf{T} \\ -\theta_{D+1}(0)\boldsymbol{\theta}(0)^\mathsf{T} \end{bmatrix} = \begin{bmatrix} \mathbf{I} \\ \mathbf{0}^\mathsf{T} \end{bmatrix} - \tilde{\boldsymbol{\theta}}(0)\boldsymbol{\theta}(0)^\mathsf{T}.$

Finally, we note that $\|\tilde{\boldsymbol{\theta}}(t)\|_2 = 1$ if $\|\tilde{\boldsymbol{\theta}}(0)\|_2 = 1$ and $\tilde{\mathbf{v}}(t) \in T_{\tilde{\boldsymbol{\theta}}(t)}\mathcal{S}_c^D$ if $\tilde{\mathbf{v}}(0) \in T_{\tilde{\boldsymbol{\theta}}(0)}\mathcal{S}_c^D$.

Appendix D. Error analysis of Spherical HMC

Following Leimkuhler and Reich (2004), we now show that the discretization error $e_n = \|\mathbf{z}(t_n) - \mathbf{z}^{(n)}\| = \|(\boldsymbol{\theta}(t_n), \mathbf{v}(t_n)) - (\boldsymbol{\theta}^{(n)}, \mathbf{v}^{(n)})\|$ (i.e. the difference between the true solution and the numerical solution) is $\mathcal{O}(\varepsilon^3)$ locally and $\mathcal{O}(\varepsilon^2)$ globally, where ε is the discretization step size. Here, we assume that $\mathbf{f}(\boldsymbol{\theta}, \mathbf{v}) := \mathbf{v}^\mathsf{T}\boldsymbol{\Gamma}(\boldsymbol{\theta})\mathbf{v} + \mathbf{G}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta})$ is smooth; hence, $\mathbf{f}$

and its derivatives are uniformly bounded as $\mathbf{z} = (\boldsymbol{\theta}, \mathbf{v})$ evolves within finite time duration T. We expand the true solution $\mathbf{z}(t_{n+1})$ at t_n:

$$
\begin{aligned}
\mathbf{z}(t_{n+1}) &= \mathbf{z}(t_n) + \dot{\mathbf{z}}(t_n)\varepsilon + \frac{1}{2}\ddot{\mathbf{z}}(t_n)\varepsilon^2 + \mathcal{O}(\varepsilon^3) \\
&= \begin{bmatrix} \boldsymbol{\theta}(t_n) \\ \mathbf{v}(t_n) \end{bmatrix} + \begin{bmatrix} \mathbf{v}(t_n) \\ -\mathbf{f}(\boldsymbol{\theta}(t_n), \mathbf{v}(t_n)) \end{bmatrix}\varepsilon + \frac{1}{2}\begin{bmatrix} -\mathbf{f}(\boldsymbol{\theta}(t_n), \mathbf{v}(t_n)) \\ -\dot{\mathbf{f}}(\boldsymbol{\theta}(t_n), \mathbf{v}(t_n)) \end{bmatrix}\varepsilon^2 + \mathcal{O}(\varepsilon^3)
\end{aligned}
\tag{73}
$$

We first consider Spherical HMC in the Cartesian coordinate, where $\mathbf{f}(\boldsymbol{\theta}, \mathbf{v}) = \|\tilde{\mathbf{v}}\|^2\boldsymbol{\theta} + [\mathbf{I} - \boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}]\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta})$. From Equation (34) we have

$$
\begin{aligned}
\mathbf{v}^{(n+1/2)} &= \mathbf{v}^{(n)} - \frac{\varepsilon}{2}(\mathbf{I} - \boldsymbol{\theta}^{(n)}(\boldsymbol{\theta}^{(n)})^{\mathsf{T}})\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta}^{(n)}) \\
\|\tilde{\mathbf{v}}^{(n+1/2)}\|^2 &= \|\tilde{\mathbf{v}}^{(n)}\|^2 - \varepsilon(\mathbf{v}^{(n)})^{\mathsf{T}}\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta}^{(n)}) + \mathcal{O}(\varepsilon^2)
\end{aligned}
\tag{74}
$$

Now we expand Equation (35) using Taylor series as follows:

$$
\begin{aligned}
\boldsymbol{\theta}^{(n+1)} &= \boldsymbol{\theta}^{(n)}[1 - \|\tilde{\mathbf{v}}^{(n+1/2)}\|^2\varepsilon^2/2 + \mathcal{O}(\varepsilon^4)] + \mathbf{v}^{(n+1/2)}\varepsilon[1 - \|\tilde{\mathbf{v}}^{(n+1/2)}\|^2\varepsilon^2/3! + \mathcal{O}(\varepsilon^4)] \\
\mathbf{v}^{(n+3/4)} &= -\boldsymbol{\theta}^{(n)}\|\tilde{\mathbf{v}}^{(n+1/2)}\|^2\varepsilon[1 - \|\tilde{\mathbf{v}}^{(n+1/2)}\|^2\varepsilon^2/3! + \mathcal{O}(\varepsilon^4)] + \mathbf{v}^{(n+1/2)}[1 - \|\tilde{\mathbf{v}}^{(n+1/2)}\|^2\varepsilon^2/2 + \mathcal{O}(\varepsilon^4)]
\end{aligned}
$$

Substituting (74) in the above equations yields

$$
\begin{aligned}
\boldsymbol{\theta}^{(n+1)} &= \boldsymbol{\theta}^{(n)} + \mathbf{v}^{(n+1/2)}\varepsilon - \boldsymbol{\theta}^{(n)}\|\tilde{\mathbf{v}}^{(n+1/2)}\|^2\varepsilon^2/2 + \mathcal{O}(\varepsilon^3) \\
&= \boldsymbol{\theta}^{(n)} + \mathbf{v}^{(n)}\varepsilon - \frac{1}{2}\mathbf{f}(\boldsymbol{\theta}^{(n)}, \mathbf{v}^{(n)})\varepsilon^2 + \mathcal{O}(\varepsilon^3) \\
\mathbf{v}^{(n+3/4)} &= \mathbf{v}^{(n+1/2)} - \boldsymbol{\theta}^{(n)}\|\tilde{\mathbf{v}}^{(n+1/2)}\|^2\varepsilon - \mathbf{v}^{(n+1/2)}\|\tilde{\mathbf{v}}^{(n+1/2)}\|^2\varepsilon^2/2 + \mathcal{O}(\varepsilon^3) \\
&= \mathbf{v}^{(n)} - [(\mathbf{I} - \boldsymbol{\theta}^{(n)}(\boldsymbol{\theta}^{(n)})^{\mathsf{T}})\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta}^{(n)})/2 + \boldsymbol{\theta}^{(n)}\|\tilde{\mathbf{v}}^{(n)}\|^2]\varepsilon \\
&\quad + [\boldsymbol{\theta}^{(n)}(\mathbf{v}^{(n)})^{\mathsf{T}}\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta}^{(n)}) - \mathbf{v}^{(n)}\|\tilde{\mathbf{v}}^{(n)}\|^2/2]\varepsilon^2 + \mathcal{O}(\varepsilon^3)
\end{aligned}
$$

With the above results, we have

$$
\begin{aligned}
\mathbf{v}^{(n+1)} =& \mathbf{v}^{(n+3/4)} - \frac{\varepsilon}{2}(\mathbf{I} - \boldsymbol{\theta}^{(n+1)}(\boldsymbol{\theta}^{(n+1)})^{\mathsf{T}})\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta}^{(n+1)}) \\
=& \mathbf{v}^{(n)} - \mathbf{f}(\boldsymbol{\theta}^{(n)}, \mathbf{v}^{(n)})\varepsilon + [\boldsymbol{\theta}^{(n)}(\mathbf{v}^{(n)})^{\mathsf{T}}\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta}^{(n)}) - \mathbf{v}^{(n)}\|\tilde{\mathbf{v}}^{(n)}\|^2/2]\varepsilon^2 \\
& - \frac{1}{2}[(\mathbf{I} - \boldsymbol{\theta}^{(n)}(\boldsymbol{\theta}^{(n)})^{\mathsf{T}})\nabla_{\boldsymbol{\theta}}^2U(\boldsymbol{\theta}^{(n)})\mathbf{v}^{(n)} - (\boldsymbol{\theta}^{(n)}(\mathbf{v}^{(n)})^{\mathsf{T}} + \mathbf{v}^{(n)}(\boldsymbol{\theta}^{(n)})^{\mathsf{T}})\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta}^{(n)})]\varepsilon^2 + \mathcal{O}(\varepsilon^3) \\
=& \mathbf{v}^{(n)} - \mathbf{f}(\boldsymbol{\theta}^{(n)}, \mathbf{v}^{(n)})\varepsilon - \frac{1}{2}\dot{\mathbf{f}}(\boldsymbol{\theta}^{(n)}, \mathbf{v}^{(n)})\varepsilon^2 + \mathcal{O}(\varepsilon^3)
\end{aligned}
$$

where for the last equality we need to show $(\mathbf{v}^{(n)})^{\mathsf{T}} \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(n)}) = -2\frac{d}{dt}\|\tilde{\mathbf{v}}^{(n)}\|^2$. This can be proved as follows:

$$
\begin{aligned}
\frac{d}{dt}\|\tilde{\mathbf{v}}\|^2 &= \frac{d}{dt}[\|\tilde{\mathbf{v}}\|^2 + v_{D+1}^2] = 2[-\mathbf{v}^{\mathsf{T}}\mathbf{f} + v_{D+1}\dot{v}_{D+1}] \\
&= 2\left[-\mathbf{v}^{\mathsf{T}}\mathbf{f} + \left(-\frac{\dot{\boldsymbol{\theta}}^{\mathsf{T}}\mathbf{v} + \boldsymbol{\theta}^{\mathsf{T}}\dot{\mathbf{v}}}{\theta_{D+1}} + \frac{\boldsymbol{\theta}^{\mathsf{T}}\mathbf{v}}{\theta_{D+1}^2}\dot{\theta}_{D+1}\right)v_{D+1}\right] \\
&= -2\left[\left(\mathbf{v} - \frac{v_{D+1}}{\theta_{D+1}}\boldsymbol{\theta}\right)^{\mathsf{T}}\mathbf{f} + \frac{v_{D+1}}{\theta_{D+1}}\|\tilde{\mathbf{v}}\|^2\right] \\
&= -2\left[\left(\mathbf{v}^{\mathsf{T}}\boldsymbol{\theta} - \frac{v_{D+1}}{\theta_{D+1}}(\|\boldsymbol{\theta}\|^2 - 1)\right)\|\tilde{\mathbf{v}}\|^2 + \left(\mathbf{v} - \frac{v_{D+1}}{\theta_{D+1}}\boldsymbol{\theta}\right)^{\mathsf{T}}[\mathbf{I} - \boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}]\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})]\right] \\
&= -2\left[\mathbf{v}^{\mathsf{T}}\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) + \left(-\mathbf{v}^{\mathsf{T}}\boldsymbol{\theta} - \frac{v_{D+1}}{\theta_{D+1}}(1 - \|\boldsymbol{\theta}\|^2)\right)\boldsymbol{\theta}^{\mathsf{T}}\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})\right] = -2\mathbf{v}^{\mathsf{T}}\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})
\end{aligned}
$$

Therefore we have

$$
\mathbf{z}^{(n+1)} := \begin{bmatrix} \boldsymbol{\theta}^{(n+1)} \\ \mathbf{v}^{(n+1)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\theta}^{(n)} \\ \mathbf{v}^{(n)} \end{bmatrix} + \begin{bmatrix} \mathbf{v}^{(n)} \\ -\mathbf{f}(\boldsymbol{\theta}^{(n)}, \mathbf{v}^{(n)}) \end{bmatrix}\varepsilon + \frac{1}{2}\begin{bmatrix} -\mathbf{f}(\boldsymbol{\theta}^{(n)}, \mathbf{v}^{(n)}) \\ -\dot{\mathbf{f}}(\boldsymbol{\theta}^{(n)}, \mathbf{v}^{(n)}) \end{bmatrix}\varepsilon^2 + \mathcal{O}(\varepsilon^3) \quad (75)
$$

The local error is

$$
\begin{aligned}
e_{n+1} &= \|\mathbf{z}(t_{n+1}) - \mathbf{z}^{(n+1)}\| \\
&= \left\|\begin{bmatrix} \boldsymbol{\theta}(t_n) - \boldsymbol{\theta}^{(n)} \\ \mathbf{v}(t_n) - \mathbf{v}^{(n)} \end{bmatrix} + \begin{bmatrix} \mathbf{v}(t_n) - \mathbf{v}^{(n)} \\ -[\mathbf{f}(t_n) - \mathbf{f}^{(n)}] \end{bmatrix}\varepsilon + \frac{1}{2}\begin{bmatrix} -[\mathbf{f}(t_n) - \mathbf{f}^{(n)}] \\ -[\dot{\mathbf{f}}(t_n) - \dot{\mathbf{f}}^{(n)}] \end{bmatrix}\varepsilon^2 + \mathcal{O}(\varepsilon^3)\right\| \quad (76) \\
&\leq (1 + M_1\varepsilon + M_2\varepsilon^2)e_n + \mathcal{O}(\varepsilon^3)
\end{aligned}
$$

where $M_k = c_k \sup_{t \in [0,T]} \|\nabla^k \mathbf{f}(\boldsymbol{\theta}(t), \mathbf{v}(t))\|$, $k = 1, 2$ for some constants $c_k > 0$. Accumulating the local errors by iterating the above inequality for $L = T/\varepsilon$ steps provides the following global error:

$$
\begin{aligned}
e_{L+1} &\leq (1 + M_1\varepsilon + M_2\varepsilon^2)e_L + \mathcal{O}(\varepsilon^3) \leq (1 + M_1\varepsilon + M_2\varepsilon^2)^2 e_{L-1} + 3\mathcal{O}(\varepsilon^3) \leq \cdots \\
&\leq (1 + M_1\varepsilon + M_2\varepsilon^2)^L e_1 + L\mathcal{O}(\varepsilon^3) \leq (e^{M_1 T} + T)\varepsilon^2 \to 0, \quad as\ \varepsilon \to 0
\end{aligned} \quad (77)
$$

For Spherical HMC in the spherical coordinate, we conjecture that the integrator of Algorithm 2 still has order 3 local error and order 2 global error. One can follow the same argument as above to verify this.

Appendix E. Bounce in diamond: Wall HMC for 1-norm constraint

Neal (2011) discusses the *Wall* HMC method for ∞-norm constraint only. We can however derive a similar approach for 1-norm constraint. As shown in the left panel of Figure 13, given the current state $\boldsymbol{\theta}_0$, HMC makes a proposal $\boldsymbol{\theta}$. It will hit the boundary to move from $\boldsymbol{\theta}_0$ towards $\boldsymbol{\theta}$. To determine the hit point 'X', we are required to solve for $t \in (0, 1)$ such that

$$
\|\boldsymbol{\theta}_0 + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)t\|_1 = \sum_{d=1}^{D} |\theta_0^d + (\theta^d - \theta_0^d)t| = 1 \quad (78)
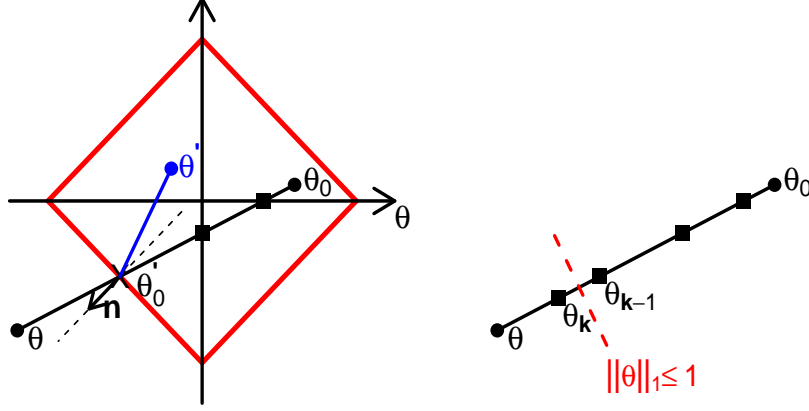$$

Figure 13: Wall HMC bounces in the 1-norm constraint domain. Left: given the current state $\boldsymbol{\theta}_0$, Wall HMC proposes $\boldsymbol{\theta}$, but bounces of the boundary and reaches $\boldsymbol{\theta}'$ instead. Right: determining the hitting time by monitoring the first intersection point with coordinate that violates the constraint.

One can find the hitting time using the bisection method. However, a more efficient method is to find the orthant in which the sampler hits the boundary, i.e., find the normal direction $\mathbf{n}$ with elements being ± 1. Then, we can find t,

$$\|\boldsymbol{\theta}_0 + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)t\|_1 = \mathbf{n}^\mathsf{T}[\boldsymbol{\theta}_0 + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)t] = 1 \implies t^* = \frac{1 - \mathbf{n}^\mathsf{T}\boldsymbol{\theta}_0}{\mathbf{n}^\mathsf{T}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)} \tag{79}$$

Therefore the hit point is $\boldsymbol{\theta}_0' = \boldsymbol{\theta}_0 + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)t^*$ and consequently the reflection point is

$$\boldsymbol{\theta}' = \boldsymbol{\theta} - 2\mathbf{n}^*\langle \mathbf{n}^*, \boldsymbol{\theta} - \boldsymbol{\theta}_0'\rangle = \boldsymbol{\theta} - 2\mathbf{n}(\mathbf{n}^\mathsf{T}\boldsymbol{\theta} - 1)/D \tag{80}$$

where $\mathbf{n}^* := \mathbf{n}/\|\mathbf{n}\|_2$ and $\mathbf{n}^\mathsf{T}\boldsymbol{\theta}_0' = 1$ because $\boldsymbol{\theta}_0'$ is on the boundary with the normal direction $\mathbf{n}^*$.

It is in general difficult to directly determine the intersection of $\boldsymbol{\theta} - \boldsymbol{\theta}_0$ with boundary. Instead, we can find its intersections with coordinate planes $\{\boldsymbol{\pi}_d\}_{d=1}^D$, where $\boldsymbol{\pi}_d := \{\boldsymbol{\theta} \in \mathbb{R}^D | \theta^d = 0\}$. The intersection times are defined as $\mathbf{T} = \{\theta_0^d/(\theta_0^d - \theta^d) | \theta_0^d \neq \theta^d\}$. We keep those between 0 and 1 and sort them in ascending order (Figure 13, right panel). Then, we find the intersection points $\{\boldsymbol{\theta}_k := \boldsymbol{\theta}_0 + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)T_k\}$ that violate the constraint $\|\boldsymbol{\theta}\| \leq 1$. Denote the first intersection point outside the constrained domain as $\boldsymbol{\theta}_k$. The signs of $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}_{k-1}$ determine the orthant of the hitting point $\boldsymbol{\theta}_0'$.

Note, for each $d \in \{1, \cdots D\}$, $(\mathrm{sign}(\boldsymbol{\theta}_k^d), \mathrm{sign}(\boldsymbol{\theta}_{k-1}^d))$ cannot be $(+, -)$ or $(-, +)$, otherwise there exists an intersection point $\boldsymbol{\theta}^* := \boldsymbol{\theta}_0 + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)T^*$ with some coordinate plane $\boldsymbol{\pi}_{d^*}$ between $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}_{k-1}$. Then $T_{k-1} < T^* < T_k$ contradicts the order of $\mathbf{T}$.

Algorithm 3 Wall HMC for 1-norm constraint (Wall HMC)

Initialize $\boldsymbol{\theta}^{(1)}$ at the current state $\boldsymbol{\theta}$ after transformation
Sample a new velocity value $\mathbf{v}^{(1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$
Calculate $H(\boldsymbol{\theta}^{(1)}, \mathbf{v}^{(1)}) = U(\boldsymbol{\theta}^{(1)}) + K(\mathbf{v}^{(1)})$
for $\ell = 1$ to L **do**
  $\mathbf{v}^{(\ell+\frac{1}{2})} = \mathbf{v}^{(\ell)} - \frac{\varepsilon}{2}\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta}^{(\ell)})$
  $\boldsymbol{\theta}^{(\ell+1)} = \boldsymbol{\theta}^{(\ell)} + \varepsilon\mathbf{v}^{(\ell+\frac{1}{2})}$
  set hit $\leftarrow$ false
  while $\|\boldsymbol{\theta}^{(\ell)}\| > 1$ **do**
    find coordinate plane intersecting times: $\mathbf{T} = \{T_d := \theta_d^{(\ell)}/(\theta_d^{(\ell)} - \theta_d^{(\ell+1)}) | \theta_d^{(\ell)} \neq \theta_d^{(\ell+1)}\}$
    sort those between 0 and 1 in ascending order: $\mathbf{T} = \{0 \leq T_k \uparrow \leq 1\}$
    find the first point in $\{\boldsymbol{\theta}_k := \boldsymbol{\theta}^{(\ell)} + (\boldsymbol{\theta}^{(\ell+1)} - \boldsymbol{\theta}^{(\ell)})T_k\}$ that violates $\|\boldsymbol{\theta}\| \leq 1$ and
denote it as $\boldsymbol{\theta}_k$
    set normal direction as $\mathbf{n} = \text{sign}(\text{sign}(\boldsymbol{\theta}_k) + \text{sign}(\boldsymbol{\theta}_{k-1}))$
    find the wall hitting time $t^* = (1 - \mathbf{n}^\mathsf{T}\boldsymbol{\theta}^{(\ell)})/(\mathbf{n}^\mathsf{T}(\boldsymbol{\theta}^{(\ell+1)} - \boldsymbol{\theta}^{(\ell)}))$
    $\boldsymbol{\theta}^{(\ell)} \leftarrow \boldsymbol{\theta}^{(\ell)} + (\boldsymbol{\theta}^{(\ell+1)} - \boldsymbol{\theta}^{(\ell)})t^*$ and $\boldsymbol{\theta}^{(\ell+1)} \leftarrow \boldsymbol{\theta}^{(\ell+1)} - 2\mathbf{n}\langle\mathbf{n}, \boldsymbol{\theta}^{(\ell+1)} - \boldsymbol{\theta}^{(\ell)}\rangle/\|\mathbf{n}\|_2^2$
    set hit $\leftarrow$ true
  end while
  if hit **then**
    $\mathbf{v}^{(\ell+\frac{1}{2})} \leftarrow (\boldsymbol{\theta}^{(\ell+1)} - \boldsymbol{\theta}^{(\ell)})\|\mathbf{v}^{(\ell+\frac{1}{2})}\|/\|\boldsymbol{\theta}^{(\ell+1)} - \boldsymbol{\theta}^{(\ell)}\|$
  end if
  $\mathbf{v}^{(\ell+1)} = \mathbf{v}^{(\ell+\frac{1}{2})} - \frac{\varepsilon}{2}\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta}^{(\ell+1)})$
end for
Calculate $H(\boldsymbol{\theta}^{(L+1)}, \mathbf{v}^{(L+1)}) = U(\boldsymbol{\theta}^{(L+1)}) + K(\mathbf{v}^{(L+1)})$
Calculate the acceptance probability $\alpha = \min\{1, \exp[-H(\boldsymbol{\theta}^{(L+1)}, \mathbf{v}^{(L+1)}) + H(\boldsymbol{\theta}^{(1)}, \mathbf{v}^{(1)})]\}$
Accept or reject the proposal according to α for the next state $\boldsymbol{\theta}'$

---

[4] Therefore any point (including $\boldsymbol{\theta}_0'$) between $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}_{k-1}$ must have the same sign as $\text{sign}(\text{sign}(\boldsymbol{\theta}_k) + \text{sign}(\boldsymbol{\theta}_{k-1}))$; that is

$$\mathbf{n} = \text{sign}(\text{sign}(\boldsymbol{\theta}_k) + \text{sign}(\boldsymbol{\theta}_{k-1})) \tag{81}$$

After moving from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$, we examine whether $\boldsymbol{\theta}'$ satisfies the constraint. If it does not satisfy the constraint, we repeat above procedure with $\boldsymbol{\theta}_0 \leftarrow \boldsymbol{\theta}_0'$ and $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}'$ until the final state is inside the constrained domain. Then we adjust the velocity direction by

$$\mathbf{v} \leftarrow (\boldsymbol{\theta}' - \boldsymbol{\theta}_0')\frac{\|\mathbf{v}\|}{\|\boldsymbol{\theta}' - \boldsymbol{\theta}_0'\|} \tag{82}$$

Algorithm 3 summarizes the above steps.

---

4. The same argument applies when $T_k = 1$, i.e. $\boldsymbol{\theta}$ is the first point outside the domain among $\{\boldsymbol{\theta}_k\}$.

References

Y. Ahmadian, J. W. Pillow, and L. Paninski. Efficient Markov Chain Monte Carlo methods for decoding neural spike trains. *Neural Computation*, 23(1):46–96, 2011.

S. Ahn, Y. Chen, and M. Welling. Distributed and adaptive darting Monte Carlo through regenerations. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AI Stat)*, 2013.

S. Ahn, B. Shahbaba, and M. Welling. Distributed Stochastic Gradient MCMC. In *International Conference on Machine Learning*, 2014.

S. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical monographs*. Oxford University Press, 2000.

C. Andrieu and E. Moulines. On the ergodicity properties of some adaptive mcmc algorithms. *Annals of Applied Probability*, 16(3):1462–1505, 2006.

M. J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, University College London, London, UK, 2003.

Alexandros Beskos, Frank J Pinski, Jesús Marıa Sanz-Serna, and Andrew M Stuart. Hybrid monte carlo on hilbert spaces. *Stochastic Processes and their Applications*, 121(10):2201–2230, 2011.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

A. E. Brockwell. Parallel markov chain monte carlo simulation by Pre-Fetching. *Journal of Computational and Graphical Statistics*, pages 246–261, 2006.

Marcus A. Brubaker, Mathieu Salzmann, and Raquel Urtasun. A family of mcmc methods on implicitly defined manifolds. In Neil D. Lawrence and Mark A. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS-12)*, volume 22, pages 161–172, 2012.

S. Byrne and M. Girolami. Geodesic Monte Carlo on Embedded Manifolds. *ArXiv e-prints*, January 2013.

B. Calderhead and M. Sustik. Sparse approximate manifolds for differential geometric mcmc. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2888–2896. 2012.

Olivier Cappé, Randal Douc, Arnaud Guillin, Jean-Michel Marin, and Christian P. Robert. Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18 (4):447–459, 2008.

R. V. Craiu, Jeffrey R., and Chao Y. Learn from thy neighbor: Parallel-chain and regional adaptive mcmc. *Journal of the American Statistical Association*, 104(488):1454–1466, 2009.

N. de Freitas, P. Højen-Sørensen, M. Jordan, and R. Stuart. Variational MCMC. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, pages 120–127, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-800-1.

Manfredo P. do Carmo. *Riemannian Geometry*. Birkhäuser Boston, 1 edition, January 1992. ISBN 0817634908.

S. Duane, A. D. Kennedy, B J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216 – 222, 1987.

Ildiko E. Frank and Jerome H. Friedman. A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, 35(2):109–135, 1993.

A. Gelfand, L. van der Maaten, Y. Chen, and M. Welling. On herding and the cycling perceptron theorem. In *Advances in Neural Information Processing Systems 23*, pages 694–702, 2010.

C. J. Geyer. Practical Markov Chain Monte Carlo. *Statistical Science*, 7(4):473–483, 1992.

Walter R. Gilks, Gareth O. Roberts, and Sujit K. Sahu. Adaptive markov chain monte carlo through regeneration. *Journal of the American Statistical Association*, 93(443):pp. 1045–1054, 1998. ISSN 01621459.

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

Gene H. Golub and Charles F. Van Loan. *Matrix computations (3rd ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996. ISBN 0-8018-5414-8.

Chris Hans. Bayesian lasso regression. *Biometrika*, 96(4):835–845, 2009.

M. Hoffman and A. Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. arxiv.org/abs/1111.4246, 2011.

Matthew Hoffman, Francis R Bach, and David M Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864, 2010.

K. Kurihara, M. Welling, and N. Vlassis. Accelerated variational Dirichlet process mixtures. In *Advances of Neural Information Processing Systems – NIPS*, volume 19, 2006.

Shiwei Lan, Vasileios Stathopoulos, Babak Shahbaba, and Mark Girolami. Markov chain monte carlo from lagrangian dynamics. *Journal of Computational and Graphical Statistics*, (just-accepted):00–00, 2014.

B. Leimkuhler and S. Reich. *Simulating Hamiltonian Dynamics*. Cambridge University Press, 2004.

J. Møller, A. Pettitt, K. Berthelsen, and R. Reeves. An efficient Markov chain Monte Carlo method for distributions with intractable normalisation constants. *Biometrica*, 93, 2006. to appear.

Iain Murray, Ryan Prescott Adams, and David J.C. MacKay. Elliptical slice sampling. *JMLR: W&CP*, 9:541–548, 2010.

Per Mykland, Luke Tierney, and Bin Yu. Regeneration in markov chain samplers. *Journal of the American Statistical Association*, 90(429):pp. 233–241, 1995. ISSN 01621459.

Peter Neal and Gareth O. Roberts. Optimal scaling for random walk metropolis on spherically constrained target densities. *Methodology and Computing in Applied Probability*, Vol.10(No.2):277–297, June 2008.

Peter Neal, Gareth O. Roberts, and Wai Kong Yuen. Optimal scaling of random walk metropolis algorithms with discontinuous target densities. *Annals of Applied Probability*, Volume 22(Number 5):1880–1927, 2012.

R. M. Neal. *Probabilistic Inference Using Markov Chain Monte Carlo Methods*. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, 1993.

R. M. Neal. The short-cut metropolis method. Technical Report 0506, Department of Statistics, University of Toronto, 2005.

R. M. Neal. MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X. L. Meng, editors, *Handbook of Markov Chain Monte Carlo*, pages 113–162. Chapman and Hall/CRC, 2011.

Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996. ISBN 0387947248.

Radford M. Neal. Slice sampling. *Annals of Statistics*, 31(3):705–767, 2003.

A. Pakman and L. Paninski. Exact Hamiltonian Monte Carlo for Truncated Multivariate Gaussians. *ArXiv e-prints*, August 2013.

Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Sam Patterson and Yee Whye Teh. Stochastic gradient riemannian langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems*, pages 3102–3110, 2013.

J. G. Propp and D. B. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. volume 9, pages 223–252, 1996.

D. Randal and Christian R. P. A vanilla rao-blackwellization of metropolis-hastings algorithms. *Annals of Statistics*, 39(1):261–277, 2011.

D. Randal, G. Arnaud, M. Jean-Michel, and R. P. Christian. Minimum variance importance sampling via population monte carlo. *ESAIM: Probability and Statistics*, 11:427–447, 2007.

G. O. Roberts and S. K. Sahu. Updating Schemes, Correlation Structure, Blocking and Parameterisation for the Gibbs Sampler. *Journal of the Royal Statistical Society, Series B*, 59:291–317, 1997.

Babak Shahbaba, Shiwei Lan, Wesley O. Johnson, and Radford M. Neal. Split hamiltonian monte carlo. *Statistics and Computing*, 24(3):339–349, 2014.

Chris Sherlock and Gareth O. Roberts. Optimal scaling of the random walk metropolis on elliptically symmetric unimodal targets. *Bernoulli*, Vol.15(No.3):774–798, August 2009.

Michael Spivak. *A Comprehensive Introduction to Differential Geometry*, volume 1. Publish or Perish, Inc., Houston, second edition, 1979.

Yee W Teh, David Newman, and Max Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *Advances in neural information processing systems*, pages 1353–1360, 2006.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1105–1112. ACM, 2009.

G. R. Warnes. The normal kernel coupler: An adaptive Markov Chain Monte Carlo method for efficiently sampling from multi-modal distributions. Technical Report Technical Report No. 395, University of Washington, 2001.

M. Welling. Herding dynamic weights to learn. In *Proc. of Intl. Conf. on Machine Learning*, 2009.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 681–688, 2011.

M. West. On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648, 1987.

Yichuan Zhang and Charles Sutton. Quasi-Newton Methods for Markov Chain Monte Carlo. In J. Shawe-Taylor, R. S. Zemel, P. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2393–2401. 2011.