Spherical Hamiltonian Monte Carlo for Constrained Target Distributions

Shiwei Lan SLAN@UCI.EDU
Department of Statistics, University of California, Irvine, CA 92697, USA.

Bo Zhou BZHOU1@UCI.EDU
Department of Statistics, University of California, Irvine, CA 92697, USA.

Babak Shahbaba BABAKS@UCI.EDU
Department of Statistics, 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 models, many copula models, and Latent Dirichlet Allocation (LDA) models. Bayesian inference involving probability distributions confined to constrained domains could be quite challenging for commonly used sampling algorithms. For such problems, we propose a novel Markov Chain Monte Carlo (MCMC) method that provides a general and computationally efficient framework for handling boundary conditions. Our method first maps the D-dimensional constrained domain of parameters to the unit ball $\mathbf{B}_0^D(1)$, then augments it to a D-dimensional sphere $\mathbf{S}^D$ such that the original boundary corresponds to the equator of $\mathbf{S}^D$. This way, our method handles the constraints implicitly by moving freely on the sphere generating proposals that remain within boundaries when mapped back to the original space.

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 models, many copula models, and Latent Dirichlet Allocation (LDA) models. Very often, the resulting models are intractable, simulating samples for Monte Carlo estimations is quite challenging (Neal & Roberts, 2008; Sherlock & Roberts, 2009; Neal et al., 2012; Brubaker et al., 2012; Pakman & Paninski, 2012), and mapping the domain to the entire Euclidean space for convenience would be computationally inefficient. In this paper, we propose a novel Markov Chain Monte Carlo (MCMC) method, which provides a natural and computationally efficient framework for sampling from constrained target distributions. Our method is based on Hamiltonian Monte Carlo (HMC) (Duane et al., 1987; Neal, 2010), which is a Metropolis algorithm with proposals guided by Hamiltonian dynamics.

In recent years, several methods have been proposed to improve the computational efficiency of HMC (Beskos et al., 2011; Girolami & Calderhead, 2011; Hoffman & Gelman, 2011; Shahbaba et al., 2013b; Lan et al., 2012; Byrne & Girolami, 2013). In general, these methods do not directly address problems with constrained target distributions. In this current paper, we focus on improving HMC-based algorithms when the target distribution is constrained by inequalities. When dealing with such constrained target distributions, the standard HMC algorithm needs to evaluate each proposal to ensure it is within the boundaries imposed by the constraints. Computationally, this is quite inefficient. Alternatively, as discussed by Neal (Neal, 2010), 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 approach, however, is not very efficient either. Byrne and Girolami (Byrne & Girolami, 2013) discuss this method for situations where constrained domains can be identified as submanifolds. Pakman and Paninski (Pakman & Paninski, 2012) also follow this idea and propose an exact HMC algorithm specifically for truncated Gaussian distributions with non-holonomic constraints. Brubaker et al. (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

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

types of constraints. In contrast, our proposed method in this paper provides a general and computationally efficient framework for handling constraints given by inequalities involving general vector norms.

In what follows, before we present our method, we provide a brief overview of HMC (Section 2). We then present our method for distributions confined to the unit ball in Section 3. The unit ball is a special case of q-norm constraints. In Section 4, we discuss the application of our method for q-norm constraints in general. In Section 5, we evaluate our proposed method using simulated and real data. Finally, we discuss future directions in Section 6.

2. HMC

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

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

$$H(\theta, p) = U(\theta) + K(p)$$

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

$$\begin{array}{rcccc} \dot{\theta} & = & \nabla_p H(\theta, p) & = & M^{-1}p \\ \dot{p} & = & -\nabla_\theta H(\theta, p) & = & -\nabla_\theta U(\theta) \end{array}$$

Note that since momentum is mass times velocity, $v = M^{-1}p$ is regarded as velocity. Throughout this paper, we express the kinetic energy K in terms of velocity, v, instead of momentum, p (Beskos et al., 2011; Lan et al., 2012).

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, 2010). We numerically 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, 2010, for more discussions).

Although HMC explores the target distribution more efficiently than random walk Metropolis, it does not fully exploit its geometric properties. To address this issue, Girolami and Calderhead (Girolami & Calderhead, 2011) propose Riemannian Manifold HMC (RMHMC), which adapts to the local Riemannian geometry of the target distribution by using a position-specific mass matrix $\boldsymbol{M} = \boldsymbol{G}(\boldsymbol{\theta})$. More specifically, they set $\boldsymbol{G}(\boldsymbol{\theta})$ to the Fisher information matrix. Our proposed sampling method can be viewed as an extension of this approach since it explores the geometry of sphere.

3. Sampling from distributions defined on the unit ball

In many cases, bounded connected constrained regions can be bijectively mapped to the D-dimensional unit ball $\mathbf{B}_0^D(1) := \{\theta \in \mathbb{R}^D : \|\theta\|_2 = \sqrt{\sum_{i=1}^D \theta_i^2} \leq 1\}$. Therefore, in this section, we first focus on distributions confined to the unit ball with the constraint $\|\theta\|_2 \leq 1$.

We start by augmenting the original D-dimensional parameter θ with an extra auxiliary variable θ_{D+1} to form an extended $(D+1)$-dimensional parameter $\tilde{\theta} = (\theta, \theta_{D+1})$ such that $\|\tilde{\theta}\|_2 = 1$ so $\theta_{D+1} = \pm\sqrt{1 - \|\theta\|_2^2}$. This way, the domain of the target distribution is changed from the unit ball $\mathbf{B}_0^D(1)$ to the D-dimensional sphere, $\mathbf{S}^D := \{\tilde{\theta} \in \mathbb{R}^{D+1} : \|\tilde{\theta}\|_2 = 1\}$, through the following transformation:

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

Note that although θ_{D+1} can be either positive or negative, its sign does not affect our Monte Carlo estimates since after applying the above transformation, we need adjust our estimates according to the change of variable theorem as follows:

$$\int_{\mathbf{B}_0^D(1)} f(\theta) d\theta_{\mathbf{B}} = \int_{\mathbf{S}_+^D} f(\tilde{\theta}) \left| \frac{d\theta_{\mathbf{B}}}{d\tilde{\theta}_{\mathbf{S}}} \right| d\tilde{\theta}_{\mathbf{S}} \tag{2}$$

where $\left|\frac{d\theta_{\mathbf{B}}}{d\tilde{\theta}_{\mathbf{S}}}\right| = |\theta_{D+1}|$ as shown in Appendix A. Here, $d\theta_{\mathbf{B}}$ and $d\tilde{\theta}_{\mathbf{S}}$ are under Euclidean measure and spherical measure respectively.

Using the above transformation, we define the dynamics on the sphere. This way, the resulting HMC sampler can move freely on $\mathbf{S}^D$ while implicitly handling the constraints imposed on the original parameters. As illustrated in Figure 1, the boundary of the constraint, i.e., $\|\theta\|_2 = 1$, corresponds to the equator on the sphere $\mathbf{S}^D$. Therefore, as the sampler moves on the sphere, passing across the equator from one hemisphere to the other translates to "bouncing back" off

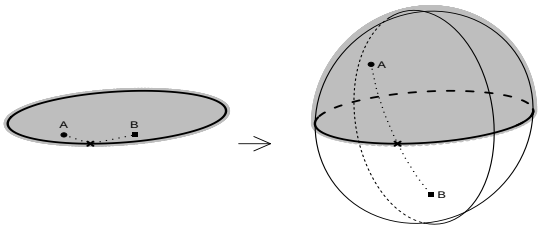

Figure 1. Transforming unit ball $\mathbf{B}_0^D(1)$ to sphere $\mathbf{S}^D$.

the the boundary in the original parameter space.

By defining HMC on the sphere, besides handling the constraints implicitly, the computational efficiency of the sampling algorithm could be improved by using splitting techniques discussed in (Beskos et al., 2011; Shahbaba et al., 2013b; Byrne & Girolami, 2013). Consider a family of target distributions, $\{f(\cdot\ ;\theta)\}$, defined on the unit ball $\mathbf{B}_0^D(1)$ (i.e., the original parameter space) endowed with the Euclidean metric $\mathbf{I}$. The potential energy is defined as $U(\theta) := -\log f(\cdot\ ;\theta)$. Associated with the auxiliary variable v (i.e., velocity), we define the kinetic energy $K(v) = \frac{1}{2}v^T\mathbf{I}v$ for $v \in T_\theta\mathbf{B}_0^D(1)$, which is a D-dimensional vector sampled from the tangent space of $\mathbf{B}_0^D(1)$. Therefore, the Hamiltonian is defined on $\mathbf{B}_0^D(1)$ as

$$H(\theta, v) = U(\theta) + K(v) = U(\theta) + \frac{1}{2}v^T\mathbf{I}v \tag{3}$$

Next, we derive the corresponding Hamiltonian function on $\mathbf{S}^D$. The potential energy $U(\tilde{\theta}) = U(\theta)$ remains the same since the distribution is fully defined in terms of the original parameter θ, i.e., the first D elements of $\tilde{\theta}$. However, the kinetic energy, $K(\tilde{v}) := \frac{1}{2}\tilde{v}^T\tilde{v}$, changes since the velocity $\tilde{v} = (v, v_{D+1})$ is now sampled from the tangent space of the sphere, $T_{\tilde{\theta}}\mathbf{S}^D := \{\tilde{v} \in \mathbb{R}^{D+1} | \tilde{\theta}^T\tilde{v} = 0\}$, with $v_{D+1} = -\theta^T v/\theta_{D+1}$. As a result, the Hamiltonian $H^*(\tilde{\theta}, \tilde{v})$ is defined on the sphere $\mathbf{S}^D$ as follows:

$$H^*(\tilde{\theta}, \tilde{v}) = U(\tilde{\theta}) + K(\tilde{v}) \tag{4}$$

Viewing $\{\theta, \mathbf{B}_0^D(1)\}$ as a coordinate chart of $\mathbf{S}^D$, this is equivalent to replacing the Euclidean metric $\mathbf{I}$ with the *canonical spherical metric* $\mathbf{G_S} = I_D + \theta\theta^T/(1 - \|\theta\|_2^2)$. Therefore, we can write the Hamiltonian function (4) as

$$H^*(\tilde{\theta}, \tilde{v}) = U(\tilde{\theta}) + \frac{1}{2}\tilde{v}^T\tilde{v} = U(\theta) + \frac{1}{2}v^T\mathbf{G_S}v \tag{5}$$

More details are provided in Appendix A.

Now we can sample the velocity $v \sim \mathcal{N}(0, \mathbf{G_S}^{-1})$ and set $\tilde{v} = \begin{bmatrix} I \\ -\theta^T/\theta_{D+1} \end{bmatrix} v$. Alternatively, we can sample $\tilde{v}$ directly from the standard $(D+1)$-dimensional Gaussian,

$$\tilde{v} \sim \mathcal{N}\left(0, \begin{bmatrix} I \\ -\theta^T/\theta_{D+1} \end{bmatrix} \mathbf{G_S}^{-1} \begin{bmatrix} I - \theta/\theta_{D+1} \end{bmatrix}\right) \tag{6}$$

which simplifies to

$$\tilde{v} \sim N(0, I_{D+1} - \tilde{\theta}\tilde{\theta}^T) \tag{7}$$

The Hamiltonian function (5) can be used to define the Hamilton dynamics on the Riemannian manifold $(\mathbf{B}_0^D(1), \mathbf{G_S})$ in terms of (θ, p), or equivalently as the following Lagrangian dynamics in terms of (θ, v) (Lan et al., 2012):

$$\begin{aligned} \dot{\theta} &= v \\ \dot{v} &= -v^T\Gamma v - \mathbf{G_S}^{-1}\nabla U(\theta) \end{aligned} \tag{8}$$

where Γ are the Christoffel symbols of second kind derived from $\mathbf{G_S}$. The Hamiltonian (5) is preserved under Lagrangian dynamics (8). (See Lan et al., 2012, for more discussion).

(Byrne & Girolami, 2013) split the Hamiltonian (5) as follows:

$$H^*(\tilde{\theta}, \tilde{v}) = U(\theta)/2 + \frac{1}{2}v^T\mathbf{G_S}v + U(\theta)/2 \tag{9}$$

However, their approach requires the manifold to be embedded in the Euclidean space. To avoid this assumption, instead of splitting the Hamilton dynamics, we split the corresponding Lagrangian dynamics (8) as follows:

$$\begin{cases} \dot{\theta} = 0 \\ \dot{v} = -\frac{1}{2}\mathbf{G_S}^{-1}\nabla U(\theta) \end{cases} \qquad \begin{cases} \dot{\theta} = v \\ \dot{v} = -v^T\Gamma v \end{cases} \tag{10}$$

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

$$\begin{aligned} \tilde{\theta}(t) &= \tilde{\theta}(0) \\ \tilde{v}(t) &= \tilde{v}(0) - \tfrac{t}{2}\left(\begin{bmatrix} I_D \\ 0 \end{bmatrix} - \tilde{\theta}(0)\theta(0)^T\right)\nabla U(\theta(0)) \end{aligned} \tag{11}$$

where t denotes time.

The second dynamics (on the right) only involves the kinetic energy; hence, it is equivalent to the geodesic flow on the sphere $\mathbf{S}^D$ with a *great circle* (orthodrome or Riemannian circle) as its analytical solution (see supplementary document at `http://www.ics.uci.edu/~slan/SphHMC` for more details),

$$\begin{aligned} \tilde{\theta}(t) &= \tilde{\theta}(0)\cos(\|\tilde{v}(0)\|_2 t) + \tfrac{\tilde{v}(0)}{\|\tilde{v}(0)\|_2}\sin(\|\tilde{v}(0)\|_2 t) \\ \tilde{v}(t) &= -\tilde{\theta}(0)\|\tilde{v}(0)\|_2\sin(\|\tilde{v}(0)\|_2 t) + \tilde{v}(0)\cos(\|\tilde{v}(0)\|_2 t) \end{aligned} \tag{12}$$

Algorithm 1 Spherical HMC

Initialize $\tilde{\theta}^{(1)}$ at current $\tilde{\theta}$ after transformation
Sample a new momentum value $\tilde{v}^{(1)} \sim \mathcal{N}(0, I_{D+1})$
Set $\tilde{v}^{(1)} \leftarrow \tilde{v}^{(1)} - \tilde{\theta}^{(1)}(\tilde{\theta}^{(1)})^T\tilde{v}^{(1)}$
Calculate $H(\tilde{\theta}^{(1)}, \tilde{v}^{(1)}) = U(\theta^{(1)}) + K(\tilde{v}^{(1)})$
for $\ell = 1$ to L **do**

$$\tilde{v}^{(\ell+\frac{1}{2})} = \tilde{v}^{(\ell)} - \frac{\epsilon}{2}\left(\begin{bmatrix} I_D \\ 0 \end{bmatrix} - \tilde{\theta}^{(\ell)}(\theta^{(\ell)})^T\right)\nabla U(\theta^{(\ell)})$$

$$\tilde{\theta}^{(\ell+1)} = \tilde{\theta}^{(\ell)}\cos(\|\tilde{v}^{(\ell+\frac{1}{2})}\|\epsilon) + \frac{\tilde{v}^{(\ell+\frac{1}{2})}}{\|\tilde{v}^{(\ell+\frac{1}{2})}\|}\sin(\|\tilde{v}^{(\ell+\frac{1}{2})}\|\epsilon)$$

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

$$\tilde{v}^{(\ell+1)} = \tilde{v}^{(\ell+\frac{1}{2})} - \frac{\epsilon}{2}\left(\begin{bmatrix} I_D \\ 0 \end{bmatrix} - \tilde{\theta}^{(\ell+1)}(\theta^{(\ell+1)})^T\right)\nabla U(\theta^{(\ell+1)})$$

end for
Calculate $H(\tilde{\theta}^{(L+1)}, \tilde{v}^{(L+1)}) = U(\theta^{(L+1)}) + K(\tilde{v}^{(L+1)})$
Calculate the acceptance probability

$$\alpha = \exp\{-H(\tilde{\theta}^{(L+1)}, \tilde{v}^{(L+1)}) + H(\tilde{\theta}^{(1)}, \tilde{v}^{(1)})\}$$

Accept or reject the proposal according to α
Calculate the corresponding weight $|\theta_{D+1}^{(n)}|$

Note that (11) and (12) are both symplectic. Due to the explicit formula for the geodesic flow on sphere, the second dynamics in (10) is simulated exactly. Therefore, updating $\tilde{\theta}$ does not involve discretization error so we can use large step sizes. This could lead to improved computational efficiency. Since this step is in fact a rotation on sphere, we set the trajectory length to be $2\pi/D$ and randomize the number of leapfrog steps to avoid periodicity. Algorithm 1 shows the steps for implementing this approach, henceforth called *Spherical HMC*.

4. Norm constraints

The unit ball region discussed in the previous section is in fact a special case of q-norm constraints. In this section we discuss q-norm constraints in general and show how they can be transformed to the unit ball so that the Spherical HMC method can be used. In general, these constraints are expressed in terms of q-norm of parameters,

$$\|\beta\|_q = \begin{cases} (\sum_{i=1}^D |\beta_i|^q)^{1/q}, & q \in (0, +\infty) \\ \max_{1\le i\le D}|\beta_i|, & q = +\infty \end{cases} \tag{13}$$

For example, when β are regression parameters, $q = 2$ corresponds to ridge regression and $q = 1$ corresponds to Lasso (Tibshirani, 1996). In what follows, we show how this type of constraints can be transformed to $\mathbf{S}^D$.

4.1. Norm constraints with $q = +\infty$

When $q = +\infty$, the distribution is confined to a hypercube. Note that hypercubes, and in general hyper-rectangles, can be transformed to the unit hypercube, $\mathbf{C}^D := [-1, 1]^D = \{\beta \in \mathbb{R}^D : \|\beta\|_\infty \le 1\}$, by proper shifting and scaling of the original parameters. (Neal, 2010) discusses this kind of constraints, which could be handled by adding a term to the energy function such that the energy goes to infinity for values that violate the constraints. This creates "energy walls" at boundaries. As a result, the sampler bounces off the energy wall whenever it reaches the boundaries. This approach, henceforth called *Wall HMC*, has limited applications and tends to be computationally inefficient.

To use Spherical HMC, the unit hypercube can be transformed to its inscribed unit ball through the following map:

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

Further, as discussed in the previous section, the resulting unit ball can be mapped to sphere $\mathbf{S}^D$ through $T_{\mathbf{B}\to\mathbf{S}}$ for which the Spherical HMC can be used. See Appendix B for the derivation of the corresponding weights needed for the change of variable.

4.2. Norm constraints with $q \in (0, +\infty)$

A domain constrained by q-norm $\mathbf{Q}^D := \{x \in \mathbb{R}^D : \|\beta\|_q \le 1\}$ for $q \in (0, +\infty)$ can be transformed to the unit ball $\mathbf{B}_0^D(1)$ via the following map:

$$T_{\mathbf{Q}\to\mathbf{B}} : \mathbf{Q}^D \to \mathbf{B}_0^D(1), \quad \beta_i \mapsto \theta_i = \mathrm{sgn}(\beta_i)|\beta_i|^{q/2} \tag{15}$$

As before, the unit ball can be transformed to the sphere for which we can use the Spherical HMC method. See Appendix B for the derivation of the corresponding weights required for the change of variable.

5. Experimental results

In this section, we evaluate our proposed methods, Spherical HMC, by comparing its efficiency to that of Random Walk Metropolis (RWM) and Wall HMC using simulated and real data. To this end, we define efficiency in terms of time-normalized effective sample size (ESS). Given B MCMC samples for each parameter, ESS = $B[1 + 2\Sigma_{k=1}^K\gamma(k)]^{-1}$, where $\Sigma_{k=1}^K\gamma(k)$ is the sum of K monotone sample autocorrelations (Geyer, 1992). We use the minimum ESS normalized by the CPU time, s (in seconds), as the overall measure of efficiency: $\min(\mathrm{ESS})/\mathrm{s}$. All computer codes are available online at `http://www.ics.uci.edu/~slan/SphHMC`.

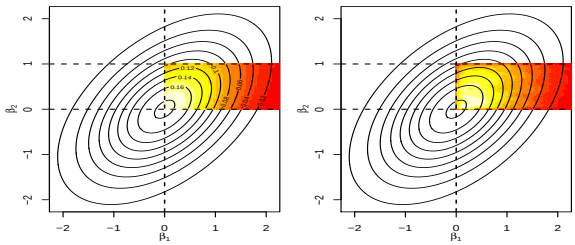

Figure 2. Density plots of a truncated bivariate Gaussian using exact density function (left) and MCMC samples from Spherical HMC (Right).

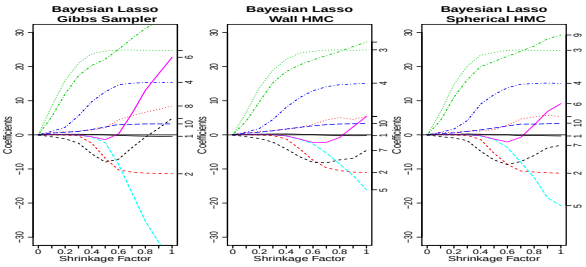

Figure 3. Bayesian Lasso using three different sampling algorithms: Gibbs sampler (left), Wall HMC (middle) and Spherical HMC (right).

Table 1. Comparing RWM, Wall HMC, and Spherical HMC in terms of acceptance probability (AP), seconds (s) per iteration, and Min(ESS)/s.

Dim	Method	AP	s/Iteration	Min(ESS)/s
D=10	RWM	0.64	1.6E-04	8.80
	Wall HMC	0.93	5.8E-04	426.79
	Spherical HMC	0.81	9.7E-04	602.78
D=100	RWM	0.72	1.3E-03	0.06
	Wall HMC	0.94	1.4E-02	14.23
	Spherical HMC	0.88	1.5E-02	40.12

5.1. Truncated Multivariate Gaussian

For illustration purposes, we first start with a truncated bivariate Gaussian distribution,

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

$$0 \le \beta_1 \le 5, \qquad 0 \le \beta_2 \le 1$$

The lower and upper limits are $l = (0, 0)$ and $u = (5, 1)$ respectively. The original rectangle domain can be mapped to the 2-dimensional unit sphere through the following transformation:

$$T : [0, 5] \times [0, 1] \to S^2,\ \beta \mapsto \beta' = (2\beta - (u + l))/(u - l)$$
$$\mapsto \theta = \beta' \frac{\|\beta'\|_\infty}{\|\beta'\|_2} \mapsto \tilde{\theta} = \left(\theta, \sqrt{1 - \|\theta\|_2^2}\right)$$

The left panel of Figure 2 shows the heatmap based on the exact density function, and the right panel shows the corresponding heatmap based on MCMC samples from Spherical HMC.

To evaluate the efficiency of the above-mentioned methods (RWM, Wall HMC, and Spherical HMC), 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 assume $0 \le \beta_i \le u_i$, where u_i (i.e., the upper bound) is set to 5 when $i = 1$; otherwise, it is set to 0.5.

For each method, we obtain 10000 MCMC samples after discarding the initial 1000 samples. We set the tuning parameters of algorithms such that their overall acceptance rates are within a reasonable range. As shown in Table 1, Spherical HMC is substantially more efficient than RWM and Wall HMC. For RWM, proposed states are rejected about 95% of times due to violation of constraints. On average, Wall HMC bounces off the wall around 7.68 and 31.10 times per iteration for $D = 10$ and $D = 100$ respectively. In contrast, by augmenting the parameter space, Spherical HMC handles the constraints in an efficient way.

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 (Tibshirani, 1996). In this approach, the coefficients are obtained by minimizing the residual sum of squares (RSS) subject to $\sum_{j=1}^{D} |\beta_j| \le t$. Park and Casella (Park & Casella, 2008) and Hans (Hans, 2009) have proposed a Bayesian alternative method, called Bayesian Lasso, where the penalty term is replaced by a prior distribution of the form $P(\beta) \propto \exp(-\lambda|\beta_j|)$, 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 method 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 use Spherical HMC with the transformation discussed in Section 4.2.

We evaluate our method based on the diabetes data set (N=442, D=10) discussed in (Park & Casella, 2008). Figure 3 compares coefficient estimates given by the Gibbs sampler (Park & Casella, 2008), Wall HMC, and Spherical HMC algorithms as the shrinkage factor $s :=$

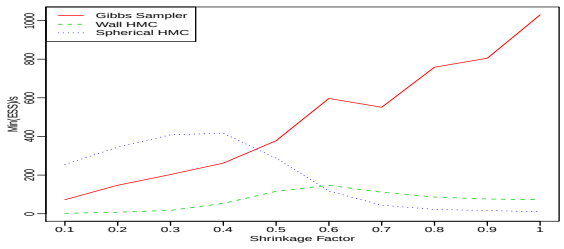

Figure 4. Sampling efficiency of different algorithms for Bayesian Lasso based on the diabetes dataset.

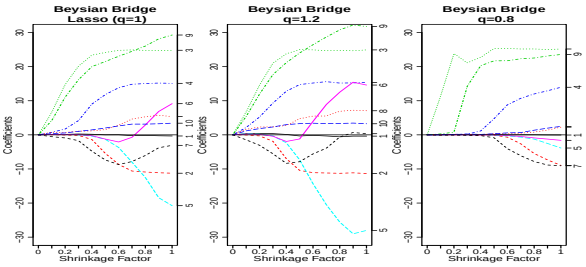

Figure 5. Bayesian Bridge Regression by Spherical HMC: Lasso (q=1, left), q=1.2 (middle), and q=0.8 (right).

$\|\hat{\beta}^{\text{Lasso}}\|_1/\|\hat{\beta}^{\text{OLS}}\|_1$ 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 4 compares the sampling efficiency of these three methods. As we impose tighter constraints (i.e., lower shrinkage factors s), our method becomes substantially more efficient than the Gibbs sampler and Wall HMC.

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 & Friedman, 1993), where the coefficients are obtained by minimizing the residual sum of squares subject to $\sum_{j=1}^{D}|\beta_j|^q \le t$. 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).

As mentioned earlier, our Spherical HMC method can easily handle this type of constraints through the following transformation:

$$T : Q^D \to S^D,\ \beta_i \mapsto \beta_i' = \beta_i/t$$
$$\mapsto \theta_i = \text{sgn}(\beta_i')|\beta_i'|^{q/2},\ \theta \mapsto \tilde{\theta} = \left(\theta, \sqrt{1-\|\theta\|_2^2}\right)$$

Figure 5 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 & 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. Modeling synchrony among multiple neurons

Shahbaba et al. (Shahbaba et al., 2013a) have recently proposed a semiparametric Bayesian model to capture dependencies among multiple neurons by detecting their cofiring patterns over time. In this approach, after discretizing time, there is at most one spike in each interval. The resulting sequence of 1's (spike) and 0's (silence) for each neuron is called a *spike train*, which is denoted as Y and is modeled using the logistic function of a continuous latent variable with a Gaussian process prior. For n neurons, the joint probability distribution of spike trains, $Y_1, \ldots, Y_n$, is coupled to the marginal distributions using a parametric copula model. Let H be n-dimensional distribution functions with marginals $F_1, ..., F_n$. In general, an n-dimensional copula is a function with the following form:

$$H(y_1, ..., y_n) = \mathcal{C}(F_1(y_1), ..., F_n(y_n)),\ \text{for all } y_1, \ldots, y_n$$

Here, $\mathcal{C}$ defines the dependence structure between the marginals. Shahbaba et al. (Shahbaba et al., 2013a) use a special case of the Farlie-Gumbel-Morgenstern (FGM) copula family (Farlie, 1960; Gumbel, 1960; Morgenstern, 1956; Nelsen, 1998), for which $\mathcal{C}$ has the following form:

$$\Big[1 + \sum_{k=2}^{n} \sum_{1\le j_1<\cdots<j_k\le n} \beta_{j_1 j_2 \ldots j_k} \prod_{l=1}^{k}(1-F_{j_l})\Big] \prod_{i=1}^{n} F_i$$

where $F_i = F_i(y_i)$. Restricting the model to second-order interactions, we have

$$H(y_1, \ldots, y_n) = \Big[1 + \sum_{1\le j_1<j_2\le n} \beta_{j_1 j_2} \prod_{l=1}^{2}(1-F_{j_l})\Big] \prod_{i=1}^{n} F_i$$

Here, $F_i = P(Y_i \le y_i)$ for the ith neuron ($i = 1, \ldots, n$), where $y_1, \ldots, y_n$ denote the firing status of n neurons at time t. β_{j_1,j_2} captures the relationship between the j_1th and j_2th neurons, with $\beta_{j_1,j_2} = 0$ interpreted as "no relationship" between the two neurons. To ensure that probability distribution functions remain within $[0, 1]$, the following constraints on all $\binom{n}{2}$ parameters, $\beta_{j_1 j_2}$, are imposed:

$$1 + \sum_{1\le j_1<j_2\le n} \beta_{j_1 j_2} \prod_{l=1}^{2} \epsilon_{j_l} \ge 0, \quad \epsilon_1, \cdots, \epsilon_n \in \{-1, 1\}$$

Table 2. Comparing RWM, Wall HMC, and Spherical HMC based on the copula model.

Scenario	Method	AP	s/Iteration	Min(ESS)/s
I	RWM	0.69	8.2	2.8E-04
	Wall HMC	0.67	17.0	7.0E-03
	Spherical HMC	0.83	17.0	2.0E-02
II	RWM	0.67	8.1	2.8E-04
	Wall HMC	0.75	19.4	1.8E-03
	Spherical HMC	0.81	18.0	2.2E-02

Considering all possible combinations of ϵ_{j_1} and ϵ_{j_2} in the above condition, there are $n(n-1)$ linear inequalities, which can be expressed as $\sum_{1\le j_1<j_2\le n}|\beta_{j_1 j_2}|\le 1$. For this model, we can use the square root mapping described in section 4.2 to transform the original domain ($q=1$) of parameters to the unit ball before using Spherical HMC.

We apply our method to a real dataset based on an experiment investigating the role of prefrontal cortical area in rats with respect to reward-seeking behavior discussed in (Shahbaba et al., 2013a). Here, we focus on 5 simultaneously recorded neurons under two scenarios: I) rewarded (pressing a lever by rats delivers 0.1 ml of 15% sucrose solution), and II) non-rewarded (nothing happens after pressing a lever by rats). There are 51 trails for each scenario. The copula model detected significant associations among three neurons: the first and forth neurons ($\beta_{1,4}$) under the rewarded scenario, and the third and forth neurons ($\beta_{3,4}$) under the non-rewarded scenario. All other parameters were deemed non-significant (based on 95% posterior probability intervals). As we can see in Table 2, Spherical HMC is order(s) of magnitudes more efficient than RWM and Wall HMC.

6. Discussion

We have introduced a new efficient sampling algorithm for constrained distributions. Our method first maps the parameter space to the unit ball and then augments the resulting space to a sphere. Further, by using the splitting strategy, we could improve the computational efficiency of our algorithm. A dynamical system is then defined on the sphere to propose new states that are guaranteed to remain within the boundaries imposed by the constraints.

In this paper, we assumed the Euclidean metric $\mathbf{I}$ on unit ball, $\mathbf{B}_0^D(1)$. The proposed approach can be extended to more complex metrics, such as the Fisher information metric $\mathbf{G_F}$, in order to exploit the geometric properties of the parameter space (Girolami & Calderhead, 2011). This way, the metric for the augmented space could be defined as $\mathbf{G_F}+\theta\theta^T/\theta_{D+1}^2$. Under such a metric however, we might not be able to find the geodesic flow analytically. This could undermine the added benefit from using the Fisher information metric.

Acknowledgments

We would like to thank Jeffrey Streets, Max Welling, and Alexander Ihler for helpful discussion. This work is supported by NSF grant IIS-1216045 and NIH grant R01-AI107034.

Appendix

A. From unit ball to sphere

Consider the D-dimensional ball $\mathbf{B}_0^D(1)=\{\theta\in\mathbb{R}^D:\|\theta\|_2\le 1\}$ and the D-dimensional sphere $\mathbf{S}^D=\{\tilde\theta=(\theta,\theta_{D+1})\in\mathbb{R}^{D+1}:\|\tilde\theta\|_2=1\}$. Note that $\{\theta,\mathbf{B}_0^D(1)\}$ can be viewed as a coordinate chart for $\mathbf{S}^D$. The first fundamental form ds^2 (i.e., squared infinitesimal length of a curve) for $\mathbf{S}^D$ is explicitly expressed in terms of the differential form $d\theta$ and the *canonical metric* $\mathbf{G_S}$ as

$$ds^2=\langle d\theta,d\theta\rangle_{\mathbf{G_S}}=d\theta^T\mathbf{G_S}d\theta$$

which can be obtained as follows (Spivak, 1979):

$$\begin{aligned}ds^2&=\sum_{i=1}^{D+1}d\theta_i^2=\sum_{i=1}^{D}d\theta_i^2+(d(\theta_{D+1}(\theta)))^2\\&=d\theta^Td\theta+\frac{(\theta^Td\theta)^2}{1-\|\theta\|_2^2}=d\theta^T[I+\theta\theta^T/\theta_{D+1}^2]d\theta\end{aligned}$$

Therefore, the canonical metric $\mathbf{G_S}$ of $\mathbf{S}^D$ is

$$\mathbf{G_S}=I_D+\frac{\theta\theta^T}{\theta_{D+1}^2}$$

For any vector $\tilde v=(v,v_{D+1})\in T_{\tilde\theta}\mathbf{S}^D=\{\tilde v\in\mathbb{R}^{D+1}:\tilde\theta^T\tilde v=0\}$, one could view $\mathbf{G_S}$ as a mean to express the length of $\tilde v$ in v:

$$\begin{aligned}v^T\mathbf{G_S}v&=\|v\|_2^2+\frac{v^T\theta\theta^Tv}{\theta_{D+1}^2}=\|v\|_2^2+\frac{(-\theta_{D+1}v_{D+1})^2}{\theta_{D+1}^2}\\&=\|v\|_2^2+v_{D+1}^2=\|\tilde v\|_2^2\end{aligned}$$

The determinant of the canonical metric $\mathbf{G_S}$ is given by the matrix determinant lemma,

$$\det G_{\mathbf{S}}=\det(I_D+\frac{\theta\theta^T}{\theta_{D+1}^2})=1+\frac{\theta^T\theta}{\theta_{D+1}^2}=\frac{1}{\theta_{D+1}^2}$$

and the inverse of $\mathbf{G_S}$ is obtained by the Sherman-Morrison-Woodbury formula (Golub & Van Loan, 1996)

$$G_{\mathbf{S}}^{-1}=\left[I_D+\frac{\theta\theta^T}{\theta_{D+1}^2}\right]^{-1}=I_D-\frac{\theta\theta^T/\theta_{D+1}^2}{1+\theta^T\theta/\theta_{D+1}^2}=I_D-\theta\theta^T$$

We now find the Jacobian determinant of $T_{\mathbf{S}\to\mathbf{B}}$. Using the volume form (Spivak, 1979), we have

$$\int_{\mathbf{S}_+^D}f(\tilde\theta)d\tilde\theta_{\mathbf{S}}=\int_{\mathbf{B}_0^D(1)}f(\theta)\sqrt{\det G_{\mathbf{S}}}d\theta_{\mathbf{B}}$$

The transformation $T_{\mathbf{B}\to\mathbf{S}} : \theta \mapsto \tilde{\theta} = (\theta, \theta_{D+1} = \sqrt{1-\|\theta\|_2^2})$ bijectively maps the unit ball $\mathbf{B}_0^D(1)$ to the upper-hemisphere $\mathbf{S}_+^D$. Using the change of variable theorem, we have

$$\int_{\mathbf{S}_+^D} f(\tilde{\theta}) d\tilde{\theta}_{\mathbf{S}} = \int_{\mathbf{B}_0^D(1)} f(\theta) \left| \frac{d\tilde{\theta}_{\mathbf{S}}}{d\theta_{\mathbf{B}}} \right| d\theta_{\mathbf{B}}$$

from which we can obtain the Jacobian determinant of $T_{\mathbf{B}\to\mathbf{S}}$ as follows:

$$\left| \frac{d\tilde{\theta}_{\mathbf{S}}}{d\theta_{\mathbf{B}}} \right| = \sqrt{\det G_{\mathbf{S}}} = 1/|\theta_{D+1}|$$

Therefore, the Jacobian determinant of $T_{\mathbf{S}\to\mathbf{B}}$ is $|\theta_{D+1}|$.

B. Transformations between different constrained domains

Denote the general hyper-rectangle type constrained domain as $\mathbf{R}^D := \{\beta \in \mathbb{R}^D : l \le \beta \le u\}$. For transformations $T_{\mathbf{S}\to\mathbf{R}}$ and $T_{\mathbf{S}\to\mathbf{Q}}$, we can find the Jacobian determinants as follows. First, we note

$$\begin{aligned}
&T_{\mathbf{S}\to\mathbf{R}} = T_{\mathbf{C}\to\mathbf{R}} \circ T_{\mathbf{B}\to\mathbf{C}} \circ T_{\mathbf{S}\to\mathbf{B}} \\
&T_{\mathbf{S}\to\mathbf{B}} : \tilde{\theta} \mapsto \theta \\
&T_{\mathbf{B}\to\mathbf{C}} : \theta \mapsto \beta' = \theta \frac{\|\theta\|_2}{\|\theta\|_\infty} \\
&T_{\mathbf{C}\to\mathbf{R}} : \beta' \mapsto \beta = \frac{u-l}{2}\beta' + \frac{u+l}{2}
\end{aligned}$$

The corresponding Jacobian matrices are

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

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

$$\begin{aligned}
&|dT_{\mathbf{S}\to\mathbf{R}}| = |dT_{\mathbf{C}\to\mathbf{R}}|\,|dT_{\mathbf{B}\to\mathbf{C}}|\,|dT_{\mathbf{S}\to\mathbf{B}}| \\
&= \left| \frac{d\beta}{d(\beta')^T} \right| \left| \frac{d\beta'}{d\theta^T} \right| \left| \frac{d\theta_{\mathbf{B}}}{d\tilde{\theta}_{\mathbf{S}}} \right| = |\theta_{D+1}| \frac{\|\theta\|_2^D}{\|\theta\|_\infty^D} \prod_{i=1}^D \frac{u_i - l_i}{2}
\end{aligned}$$

Next, we note

$$T_{\mathbf{S}\to\mathbf{Q}} = T_{\mathbf{B}\to\mathbf{Q}} \circ T_{\mathbf{S}\to\mathbf{B}} : \tilde{\theta} \mapsto \theta \mapsto \beta = \mathrm{sgn}(\theta)|\theta|^{2/q}$$

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

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

Therefore the Jacobian Determinant of $T_{\mathbf{S}\to\mathbf{Q}}$ is

$$\begin{aligned}
&|dT_{\mathbf{S}\to\mathbf{Q}}| = |dT_{\mathbf{B}\to\mathbf{Q}}|\,|dT_{\mathbf{S}\to\mathbf{B}}| \\
&= \left| \frac{d\beta}{d\theta^T} \right| \left| \frac{d\theta_{\mathbf{B}}}{d\tilde{\theta}_{\mathbf{S}}} \right| = \left(\frac{2}{q}\right)^D \left(\prod_{i=1}^D |\theta_i| \right)^{2/q-1} |\theta_{D+1}|
\end{aligned}$$

C. Splitting Hamilton dynamics on $\mathbf{S}^D$

Splitting the Hamiltonian function and its usefulness in improving HMC is a well-studied topic of research (Leimkuhler & Reich, 2004; Shahbaba et al., 2013b; Byrne & Girolami, 2013). Splitting the Lagrangian function (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 & Girolami, 2013),

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

The corresponding systems of differential equations,

$$\left\{ \begin{array}{rcl} \dot{\theta} &=& 0 \\ \dot{p} &=& -\frac{1}{2}\nabla U(\theta) \end{array} \right. \qquad \left\{ \begin{array}{rcl} \dot{\theta} &=& {\mathbf{G}_{\mathbf{S}}}^{-1} p \\ \dot{p} &=& -\frac{1}{2} p^T {\mathbf{G}_{\mathbf{S}}}^{-1} d\mathbf{G}_{\mathbf{S}} {\mathbf{G}_{\mathbf{S}}}^{-1} p \end{array} \right.$$

can be written in terms of Lagrangian dynamics as follows: in (θ, v) (Lan et al., 2012):

$$\left\{ \begin{array}{rcl} \dot{\theta} &=& 0 \\ \dot{v} &=& -\frac{1}{2}\mathbf{G}_{\mathbf{S}}^{-1}\nabla U(\theta) \end{array} \right. \qquad \left\{ \begin{array}{rcl} \dot{\theta} &=& v \\ \dot{v} &=& -v^T \Gamma v \end{array} \right.$$

To solve the first dynamics, we note that

$$\begin{aligned}
\dot{\theta}_{D+1} = & -\frac{\theta^T}{\theta_{D+1}}\dot{\theta} & = 0 \\
\dot{v}_{D+1} = & -\frac{\dot{\theta}^T v + \theta^T \dot{v}}{\theta_{D+1}} + \frac{\theta^T v}{\theta_{D+1}^2}\dot{\theta}_{D+1} & = \frac{1}{2}\frac{\theta^T}{\theta_{D+1}} \mathbf{G}_{\mathbf{S}}^{-1}\nabla U(\theta)
\end{aligned}$$

Therefore, we have

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

where

$$\begin{aligned}
\begin{bmatrix} I \\ -\frac{\theta(0)^T}{\theta_{D+1}(0)} \end{bmatrix} [I - \theta(0)\theta(0)^T] &= \begin{bmatrix} I - \theta(0)\theta(0)^T \\ -\theta_{D+1}(0)\theta(0)^T \end{bmatrix} \\
&= \begin{bmatrix} I \\ 0 \end{bmatrix} - \tilde{\theta}(0)\theta(0)^T
\end{aligned}$$

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

References

Beskos, A., Pinski, F. J., Sanz-Serna, J. M., and Stuart, A. M. Hybrid Monte-Carlo on Hilbert spaces. *Stochastic Processes and their Applications*, 121:2201–2230, 2011.

Brubaker, M. A., Salzmann, M., and Urtasun, R.. A family of mcmc methods on implicitly defined manifolds. In Lawrence, N. D. and Girolami, M. A. (eds.), *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS-12)*, volume 22, pp. 161–172, 2012.

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

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

Farlie, D. J. G. The performance of some correlation coefficients for a general bivariate distribution. *Biometrika*, 47(3/4), 1960.

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

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

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

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

Gumbel, E. J. Bivariate exponential distributions. *Journal of the American Statistical Association*, 55:698–707, 1960.

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

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

Lan, S., Stathopoulos, V., Shahbaba, B., and Girolami, M. Lagrangian Dynamical Monte Carlo. arxiv.org/abs/1211.3759, 2012.

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

Morgenstern, D. Einfache beispiele zweidimensionaler verteilungen. *Mitteilungsblatt für Mathematische Statistik*, 8:234–235, 1956.

Neal, P. and Roberts, G. O. 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.

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

Neal, R. M. MCMC using Hamiltonian dynamics. In Brooks, S., Gelman, A., Jones, G., and Meng, X. L. (eds.), *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2010.

Nelsen, R. B. *An Introduction to Copulas (Lecture Notes in Statistics)*. Springer, 1 edition, 1998.

Pakman, A. and Paninski, L. Exact Hamiltonian Monte Carlo for Truncated Multivariate Gaussians. *ArXiv e-prints*, August 2012.

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

Shahbaba, B., Zhou, B., Ombao, H., Moorman, D., and Behseta, S. A semiparametric Bayesian model for neural coding. arXiv:1306.6103, 2013a.

Shahbaba, B., Lan, S., Johnson, W. O., and Neal, R. M. Split hamiltonian monte carlo. *Statistics and Computing*, pp. 1–11, 2013b. ISSN 0960-3174. doi: 10.1007/s11222-012-9373-1.

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

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

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

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