Constraint Relaxation for Bayesian Modeling with Parameter Constraints

Leo Duan, Akihiko Nishimura, David Dunson

Abstract: Prior information often takes the form of parameter constraints. Bayesian methods include such information through prior distributions having constrained support. By using posterior sampling algorithms, one can quantify uncertainty without relying on asymptotic approximations. However, outside of narrow settings, parameter constraints make it difficult to develop new prior and/or efficient posterior sampling algorithms. We first propose a general approach to adapt flexible classes of distribution onto constrained space, then for posterior estimation, we relax the parameters into the neighborhood of constrained space, using data augmentation or an approximation function. General off the shelf posterior sampling algorithms, such as Hamiltonian Monte Carlo (HMC), can then be used directly. We illustrate this approach through multiple examples involving equality and inequality constraints. While existing methods tend to rely on conjugate families or convenient reparameterization, our proposed approach frees us up to define new classes of models for constrained problems. We illustrate this through application to a variety of simulated and real datasets.

KEY WORDS: Simplex, Stiefel Manifold, Parameter Expansion

1 Introduction

It is extremely common to have prior information available on parameter contraints in statistical models. For example, one may have prior knowledge that a vector of parameters lies on the probability simplex or satisfies a particular set of inequality constraints. Other common examples include shape constraints on functions, positive semidefiniteness of matrices and orthogonality. There is a very rich literature on optimization subject to parameter contraints. One common approach is to rely on Lagrange and Karush-Kuhn-Tucker multipliers (Boyd and Vandenberghe, 2004). However, simply producing a point estimate is often insufficient, as uncertainty quantification (UQ) is a key component of most statistical analyses. Usual large sample asymptotic theory, for example showing asymptotic normality of statistical estimators, tends to break down in constrained inference problems. Instead, limiting distributions may have a complex form that needs to be rederived for each new type of constraint, and may be intractable. An appealing alternative is to rely on Bayesian methods for UQ, including the constraint through a prior distribution having restricted support,

and then applying Markov chain Monte Carlo (MCMC) to avoid the need for large sample approximations.

One of the most common strategies is reparameterization. By replacing the constrained parameters with functions of un-/less constrained parameters (usually at equal or less dimension), the constraint can be always satisfied. The transformation, if bijective, is known as 'coordinate system' in manifold embedding literature (Nash, 1954; Do Carmo, 2016). Examples include the stick-breaking construction for Dirichlet distribution on probability simplex CITES, and the polar coordinates for data on a hyper-sphere CITES. One can then directly assign prior on the less constrained parameters. Although this strategy has been successful, convenient coordinate system does not always exist; and heavy reparameterzation makes it difficult to induce prior property on the original space. Diaconis et al. (2013) provide a useful tutorial and cautious guide on this matter.

Alternatively, it is typical to rely on customized soluton for specific constraints. One popular strategy is to restrict focus to a prior and likelihood such that posterior sampling is tractable. For example, for modeling of data on Stiefel manifolds, von Mises-Fisher and matrix Bingham-von Mises-Fisher distribution (Khatri and Mardia, 1977; Hoff, 2009) are routinely used. Besides limiting consideration to specialized models, one drawback is that the tractable compution, especially posterior conjugacy, often breaks down under common modeling complication, such as matrix symmetry, hierarchical structures, etc.

For these reasons, it is appealing to consider approaches that do not rely on conjugate constrained distributions. Early work (Gelfand et al., 1992) suggested using general unconstrained distribution inside a simple truncated space, and running Gibbs sampling ignoring the constraint but only accepting the draws that fall into truncated space. Unfortunately, this method can be highly inefficient if constrained space has a small or zero measure, due to a high rejection probability. An alternative idea is to run MCMC ignoring the constraint, and then project draws from the unconstrained posterior to the appropriately constrained space. Such an approach was proposed for generalized linear models with order constraints by Dunson and Neelon (2003), extended to functional data with monotone or unimodal constraints (Gunn and Dunson, 2005), and recently modified to nonparametric regression with monotonicity (Lin and Dunson, 2014) or manifold (Lin et al., 2016) constraints. Another independent direction utilizes Hamiltonian Monte Carlo (HMC) that incorporates geometric structure with a Riemannian metric (Girolami and Calderhead, 2011), making proposals with high acceptance probability inside the constrained space. As this approach is costly in computation, a modified HMC based on geodesic flow was proposed for a few selected constrained space (Byrne and Girolami, 2013).

The goal of this article is to dramatically expand the families of constrained priors and develop simple computational strategy. We first introduce a general strategy to adapt existing distributions into constrained space. To enable efficient posterior computation, we *relax* the parameter into the neighborhood of the

constrained space. This accomodates the geometry of the constrained space, while enjoys simple and efficient sampling in common unconstrained space. This relxation produces an approximation for posterior under general constraints formed by equality and/or inequality, and can provide exact soluton for several common constrained space such as simplex and Stifel manifolds. Theoretic studies are conducted and original models are shown in simulations and data applications.

2 Constrained Relaxation Methodology

2.1 Deriving Constrained Distribution via Conditioning

Let $\theta \in \mathcal{D}$ denote the parameters of interests. The support $\mathcal{D}$ is a constrained space. The usual Bayesian approach assigns an existing prior density $\pi_{0,\mathcal{D}}(\theta)$ for θ only having support $\mathcal{D}$, which limits the available choices. On the other hand, in the un/less-constrained space $\mathcal{R} \supset \mathcal{D}$, there is a large family of distributions with well-studied properties. We denote such a distribution as $\pi_{\mathcal{R}}(\theta)$. It would be appealing to use it in a constrained subspace. In this article, we restrict focus on $\mathcal{R}$ being the subspace of p-dimensional Euclidean space, $\mathcal{R} \subset \mathbb{R}^p$, and denote the s-dimensional Lebesgue measure of region A as $\mu^s(A)$.

Intuitively, one could directly apply space truncation to $\theta \in \mathcal{D}$ on $\pi_{\mathcal{R}}(\theta)$, renormalizing by $1/\mu^s(\mathcal{D})$ to obtain the appropriate density. However, this often does not work as many constrained space has $\mu^s(\mathcal{D}) = \int_{\mathcal{D}} \pi_{\mathcal{R}}(\theta)\mu^p(d\theta) = 0$. Therefore, we need an alternative strategy. Considering a derived random variable $w = v(\theta)$, as long as $v : \mathcal{R} \to \mathbb{R}^d$ is measurable with respect to $\pi_{\mathcal{R}}$, one can obtain conditional density given $w = w_0$,

$$\pi_{\mathcal{R}}(\theta \mid w = w_0) = \frac{1}{m^s(w_0)J(v(\theta))}\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=w_0} \tag{1}$$

where $\mathbb{1}_E$ is an indicator equal 1 when condition E holds and 0 otherwise; $J(v(\theta))$ is the Jacobian equal to $\sqrt{\det D(v(\theta))' D(v(\theta))}$ with $D(v(\theta))$ the partial derivative matrix; $m^s(w_0) = \int_{v(\theta)=w_0} \frac{1}{J(v(\theta))}\pi_{\mathcal{R}}(\theta)\mu^s(d\theta)$. This conditional density exists if $J(v(\theta)) > 0$ and $m^s(w_0) \in (0, \infty)$, for some integer $s \leq p$ as the intrinsic dimension of $\mathcal{D}$. This is based on the coarea formula in Federer (2014). A more rigorous justification is deferred to the theory section.

A large family of constraints can be associated with certain function $v(\theta)$ fixed at certain value w_0, without loss of generality, we take $w_0 = \mathbf{0}$. We have $\mathcal{D} = \{\theta : v(\theta) = \mathbf{0}\}$. For example, one equality constraint $f(\theta) = 0$ can be associated with $v(\theta) = f(\theta)$=0; one inequality $f(\theta) < 0$ can be associated with $v(\theta) = |f(\theta)|_+ = \begin{cases} 0 \text{ if } f(x) \leq 0 \\ f(x) \text{ if } f(x) > 0 \end{cases}$ provided $f : \mathcal{R} \to \mathbb{R}$ is measurable. Although this imposes a restriction

on the suitable constrained space, there is a rich class within this category. Namely, many useful manifolds are embedded in $\mathbb{R}^p$ via equality constraint (Do Carmo, 2016).

Omitting normalization constant, the derived constrained density is:

$$\pi_{\mathcal{D}}(\theta) = \pi_{\mathcal{R}}(\theta \mid v(\theta) = \mathbf{0}) \propto \pi_{\mathcal{R}}(\theta)\mathbb{1}_{\theta\in\mathcal{D}}/J(v(\theta)). \tag{2}$$

Obviously, there are potentially more than one suitable $v(\theta)$'s for this derivation. Among those, if there exists $v(\theta)$ with constant $J(v(\theta))$ when $\theta \in \mathcal{D}$, it is more preferrable. As the associated density proportional to $\pi_{\mathcal{R}}(\theta)\mathbb{1}_{\theta\in\mathcal{D}}$ can be taken as the justified version of simple space truncation, as motivated above.

We now illustrate this strategy on deriving several distributions on a $(p-1)$-hypersphere, defined as $\mathcal{D} = \{\theta \in \mathbb{R}^p : \theta'\theta = 1\}$. A simple choice of function is $v(\theta) = \theta'\theta - 1$, which has $J(v(\theta)) = \|2\theta\| = 2$ when $\theta \in \mathcal{D}$. We start with a familiar location-scale distribution Gaussian distribution with diagonal covariance $\theta \in \mathrm{No}(F, I\sigma^2)$ as $\pi_{\mathcal{R}}$, where $F \in \mathcal{D}$ and I is the identity matrix. Conditioning on $v(\theta) = \theta'\theta - 1 = 0$ yields

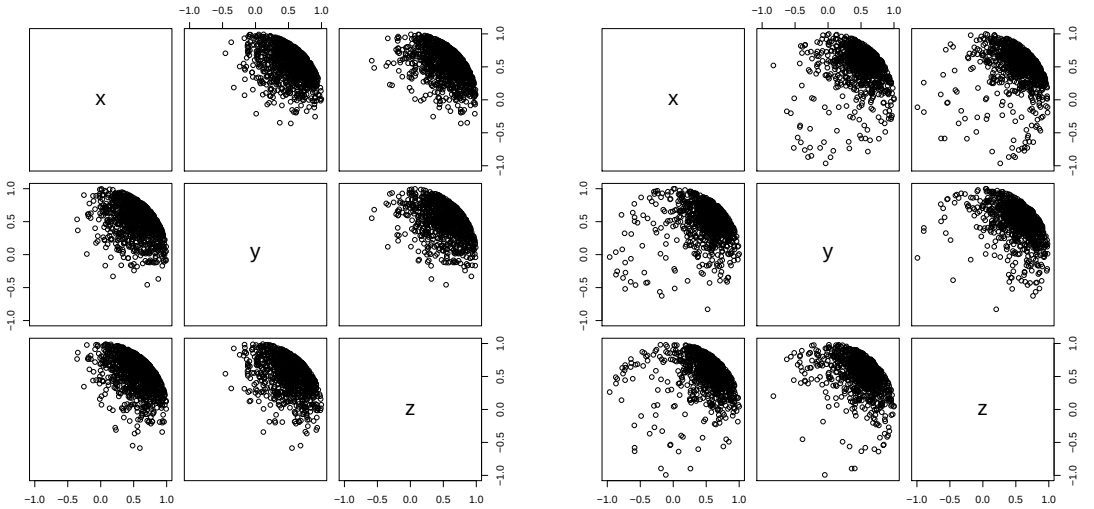
$$\begin{aligned}\pi_{\mathcal{D}}(\theta) &\propto \exp(-\frac{\|F-\theta\|^2}{2\sigma^2})\mathbb{1}_{\theta'\theta=1}/2 \\ &\propto \exp(\frac{F'}{\sigma^2}\theta)\mathbb{1}_{\theta'\theta=1},\end{aligned} \tag{3}$$

where the quadratic term $\theta'\theta$ is left out as a constant in the second line. This gives rise to the famous von Mises–Fisher distribution (Khatri and Mardia, 1977). In the Gaussian $\pi_{\mathcal{R}}(\theta)$, θ is symmetrically distributed around F, with density decaying exponentially as $\|\theta - F\|^2$ increases with rate controled by $(2\sigma^2)^{-1}$; as the constrained density $\pi_{\mathcal{D}}(\theta)$ is proportional $\pi_{\mathcal{R}}(\theta)$ on $\mathcal{D}$, it behaves similarly. Figure 1(a) shows this distribution parameterized by $F = [1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}]'$ and $\sigma^2 = 0.1$.

This strategy can be generalized to other $\pi_{\mathcal{R}}(\theta)$ based on its properties in unconstrained space $\mathcal{R}$. For example, one can start from a multivariate t-distribution $\pi_{\mathcal{R}}(\theta)$, $t_m(F, I\sigma^2)$ with m degrees of freedom, mean F and variance $I\sigma^2$, and obtain a new constrained density

$$\begin{aligned}\pi_{\mathcal{D}}(\theta) &\propto (1 + \frac{\|F-\theta\|^2}{m\sigma^2})^{-(m+p)/2}\mathbb{1}_{\theta'\theta=1}/2 \\ &\propto (1 - \frac{2F'\theta}{1+m\sigma^2+F'F})^{-(m+p)/2}\mathbb{1}_{\theta'\theta=1}\end{aligned} \tag{4}$$

As in the t-distribution, density decays polynomially as $\|F-\theta\|^2$ increases, at small m, the induced distribution (Figure 1(b) with $m = 3, p = 3$) exhibits less concentration than von Mises–Fisher on the sphere. This can be useful for robust modeling.



(a) Constrained independent Gaussian distribution (b) Constrained independent t_3 distribution

Figure 1: Sectional view of random samples from constrained distributions on a unit sphere inside $\mathbb{R}^3$. The distributions are derived through conditioning on $\theta'\theta = 1$ based on unconstrained densities of (a) $\mathrm{No}(F, \mathrm{diag}\{0.1\})$, (b) $t_3(F, \mathrm{diag}\{0.1\})$, where $F = [1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}]'$.

2.2 Constraint Relaxation for Posterior Inference

When the constrained distribution (2) is used as a prior, denoting the likelihood as $L(y;\theta)$ with y the data, one can obtain posterior:

$$\begin{aligned} \pi(\theta \mid y) &\propto L(y;\theta)\pi_{\mathcal{D}}(\theta) \\ &\propto L(y;\theta)\pi_{\mathcal{R}}(\theta)/J(v(\theta))\mathbb{1}_{\theta\in\mathcal{D}}. \end{aligned} \tag{5}$$

Clearly, the posterior also has support only inside $\mathcal{D}$ due to the inheritance of $\mathbb{1}_{\theta\in\mathcal{D}}$ from $\pi_{\mathcal{D}}(\theta)$. On the hand, the indicator is often inconvenient for posterior inference. We now present present two different strategies to relax the parameter $\theta \in \mathcal{D}$ into the neighborhood of $\mathcal{D}$, obtaining posterior θ^*. Then we directly use θ^* as approximation or project θ^* back to $\mathcal{D}$ for exact inference. We refer this strategy as **COnstraint RElaxation (CORE)**.

2.2.1 Approximation CORE

We first present a general approximation strategy. By putting positive mass around $v(\theta) = \mathbf{0}$, and let it decay exponentially, we relax the support of θ into a neighborhood surrounding $\mathcal{D}$. This generate approximate posterior sample $\theta^* \in \mathcal{R}$ via

$$\tilde{\pi}(\theta^* \mid y) = \frac{1}{m(\lambda)} L(y;\theta^*)\pi_{\mathcal{R}}(\theta^*)/J(v(\theta^*)) \exp(-\sum_{k=1}^{K} |v_k(\theta^*)|^{\alpha}/\lambda_k), \tag{6}$$

where v_k is the kth equation in $v(\theta)$, α is typically chosen as 1 or 2 to mimic Laplace or Gaussian kernel, $m(\lambda)$ is an unknown normalizing constant such that $\int_{\mathcal{R}} \tilde{\pi}(\theta^* \mid y) d\theta^* = 1$; $\lambda_k \geq 0$ is a tuning parameter controling the concentration around $v(\theta^*) = \mathbf{0}$. When $\lambda_k = 0$ for all k, (6) becomes exact; using small $\lambda_k > 0$, (6) conventional Monte Carlo approach can be carried.

We use a toy example with closed-form posterior to illustrate the effect of approximation. Consider a posterior from a sum-constrained bivariate Gaussian random vector $[\theta_1, \theta_2]' \mid y \sim \text{No}(0, I)\mathbb{1}_{\theta_1+\theta_2-1=0}$. Using $v(\theta) = \theta_1 + \theta_2 - 1 = 0$, $J(v(\theta)) = \sqrt{2}$, the exact (5) is proportional to

$$\phi(\theta_1)\phi(\theta_2)\mathbb{1}_{v(\theta)=0}$$

where $\phi(.)$ is the standard normal density. The exact posterior density has closed-form

$$\pi(\theta \mid y) = \frac{\sqrt{2}}{\sqrt{2\pi}} \exp(-\frac{(\theta_1 - \frac{1}{2})^2}{2/2})\mathbb{1}_{\theta_2=1-\theta_1}$$

corresponding to $\theta_1 \mid (\theta_1 + \theta_2 = 1) \sim \text{No}(1/2, 1/2)$, $\theta_2 \mid \theta_1 \sim \delta_{1-\theta_1}(.)$. Marginally, it is a degenerate Gaussian:

$$\begin{bmatrix}\theta_1 \\ \theta_2\end{bmatrix} \sim \text{No}_\text{d}\left(\begin{bmatrix}\frac{1}{2} \\ \frac{1}{2}\end{bmatrix}, \begin{bmatrix}\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2}\end{bmatrix}\right).$$

Using $\exp(-(\theta_1^* + \theta_2^* - 1)^2/\lambda)$ to replace $\mathbb{1}_{v(\theta)=0}$, we obtain approximation $\theta_1^* \sim \text{No}(\frac{2}{\lambda+4}, \frac{\lambda+2}{\lambda+4})$, $\theta_2^* \mid \theta_1^* \sim \text{No}(\frac{2}{\lambda+2}(1-\theta_1^*), \frac{\lambda}{\lambda+2})$. Marginally,

$$\begin{bmatrix}\theta_1^* \\ \theta_2^*\end{bmatrix} \sim \text{No}\left(\begin{bmatrix}\frac{2}{\lambda+4} \\ \frac{2}{\lambda+4}\end{bmatrix}, \begin{bmatrix}\frac{\lambda+2}{\lambda+4} & -\frac{2}{\lambda+4} \\ -\frac{2}{\lambda+4} & \frac{\lambda+2}{\lambda+4}\end{bmatrix}\right).$$

As approximation, one can sample from the non-degenerate distribution using small λ. Clearly, the approximation error decreases as λ gets smaller, and becomes exact when $\lambda \to 0$. We will formalize this notion in the theory section.

2.2.2 Exact CORE

The previous strategy is generally applicable for approximating the indicator. In some cases, when the constraint function $v(\theta)$ takes simple form, one can easily project the relaxed parameter $\theta^* \in \mathcal{R}$ back into $\mathcal{D}$ and use them to reparameterize the exact posterior density. This avoids the approximation error. Although this is more limited than the previous one, it still covers a large range of useful constraints, such as norm constraint (e.g. $v(\theta) = \|\theta\|_m - 1$, with $m = 1$ in simplex modeling, $m = 2$ in hypersphere modeling), in which the projection is simply a re-normalization.

Our goal is to use variable transformation to derive a density for $\theta^* \in \mathcal{R}$ based on $\pi_{\mathcal{D}}(\theta)$, sample θ^*, and use its projection for posterior inference. One thing we need is to make the projection invertiable. Observing

the projection is from $\{\theta^* : v(\theta^*) = w\}$ to $\{\theta : v(\theta) = \mathbf{0}\}$ using $\theta = g(\theta^*; w)$ with $w = v(\theta^*)$, one can augment the density of w to θ, so that $\theta^* = g^{-1}(\theta; w)$, leading to

$$\begin{aligned} P(\theta \in A \mid y) &= \int_{\mathcal{W}} \int_{A} \pi_{\mathcal{D}}(\theta \mid y)\pi(w)d\theta dw \\ &= \int_{\mathcal{W}} \int_{g(\theta^*;w)\in A} \pi(\theta^*, w \mid y)d\theta^* dw, \qquad \pi(\theta^*, w \mid y) = \pi_{\mathcal{D}}(g(\theta^*; w) \mid y)J(g(\theta^*; w))\pi(w), \end{aligned} \tag{7}$$

for measure over $A \subset \mathcal{D}$, where $\mathcal{W}$ is the support of w, $\int_{\mathcal{W}} \pi(w)dw = 1$, $J(g(\theta^*; w))$ is the Jacobian. Since $J(g(\theta^*; w)) = \frac{1}{J(g^{-1}(\theta;w))}$, it is easy to verify the second line can transform back to first line. Since w is independent of θ, $\pi(w)$ can take any distribution as long as its support agrees with the one of $v(\theta^*)$. This is very similar to the parameter expansion strategy used in augmenting probit regression (Liu and Wu, 1999), except they relax the variance of the latent variable from 1 to a redundant parameter, and we relax the constraint function $v(\theta)$ from $\mathbf{0}$.

To provide concrete illustration of (7), we consider a toy example $(p-1)$-simplex $\{\theta : \sum_{i=1}^{p} \theta_k = 1, \theta_i \in (0, 1) \text{ for } i = 1, \ldots, p\}$, using Dirichlet distribution $\theta \mid y \in \text{Dir}(\alpha)$

$$\pi_{\mathcal{D}}(\theta \mid y) \propto \prod_{i=1}^{p} \left(\theta_i^{\alpha-1} \mathbb{1}_{\theta_i \in (0,1)}\right) \mathbb{1}_{\sum_{i=1}^{p} \theta_i = 1}. \tag{8}$$

Relaxing simplex into space $\mathcal{R} = [0, 1]^p$, one would have $\sum_{i=1}^{p} \theta_k^* - 1 = w$. A viable projection is simply re-scaling θ^* by $1/(w + 1)$. Noticing θ_p and θ_p^* are deterministic given w and other $\{\theta_i, \theta_i^*\}_{i=1,\ldots,(p-1)}$, we drop these two. Using (7) without the integration,

$$\pi(\theta^*, w \mid y) \propto (1 - \frac{\sum_{i=1}^{p-1} \theta_i^*}{w + 1}) \prod_{i=1}^{p-1} (\frac{\theta_i^*}{(w + 1)})^{\alpha-1} (w + 1)^{-(p-1)} \mathbb{1}_{(w+1)>\sum_{i=1}^{p-1} \theta_i} \pi(w) \tag{9}$$

where $(w+1)^{-(p-1)}$ is $J(g(\theta^*; w))$. Therefore, one can sample $\theta_1^*, \ldots \theta_{p-1}^*$ and w directly for this distribution. This relaxation is particularly useful when the posterior is not conjugate on the simplex, such as hyper-Dirichlet distribution CITES.

2.3 Properties

We now present the properties of the proposed approach. We first establish that the conditioning approach yields valid probability measure.

We focus on $\mathcal{R}$ being a p-dimensional Euclidean space and the intrinsic dimension of $\mathcal{D}$, $\dim(\mathcal{D}) = d \leq p$ and is integer. When $d < p$, although the p-dimensional Lebesgue measure $\mu^p(\mathcal{D}) = 0$, it is still possible

to define a conditional probability given the event $\theta \in \mathcal{D}$ (i.e. $v(\theta) = \mathbf{0}$). We utilize the concept of *regular conditional probability* (r.c.p.) (Kolmogorov, 1950). For this article to be self-contained, we list the definition as below (a more complete review can be found in Leao Jr et al. (2004)).

Let $(X, \mathscr{A}, \mu)$ be a probability space and $(Y, \mathscr{B})$ a measurable space. A function v is measurable if $v : X \to Y$, $v^{-1}(\mathscr{B}) \in \mathscr{A}$. A r.c.p is a function $f : Y \times \mathscr{A} \to [0, 1]$ satisfying:

1. $f(y, .)$ is a measure on $(X, \mathscr{A})$ for each $y \in Y$;
2. $f(., E)$ is a measurable function on $(Y, \mathscr{B})$ for each $E \in \mathscr{A}$;
3. For each $E \in \mathscr{A}$, $F \in \mathscr{B}$, $\mu(E \cap v^{-1}(F)) = \int_F f(y, E)\mu_y(dy)$, with μ_y the induced measure on $(Y, \mathscr{B})$.

Using the previous notation, we write $f(y, E) = P(\theta \in E \mid v(\theta) = y) = \int_E \pi_{\mathcal{R}}(\theta \mid v(\theta) = y)d\theta$

Remark 1. *Assuming $J(v(\theta)) > 0$ and there is a finite and non-negtive integer s such that, for some $y \in Y$,*

$$m_s(y) = \int_{\mathbb{R}^s} \frac{\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=y}}{J(v(\theta))} d\theta \in (0, \infty), \tag{10}$$

then

$$P(E \mid v(\theta) = y) = \begin{cases} \frac{1}{m_s(y)} \int_{E \cap \mathbb{R}^s} \frac{\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=y}}{J(v(\theta))} d\theta & , \text{ if } m_s(y) \in (0, \infty) \\ \delta_{x^*}(E) \text{ with fixed } x^* \in \mathbb{R}^p & , \text{ if } m_s(y) \in \{0, \infty\} \end{cases} \tag{11}$$

is a valid r.c.p..

Proof. The first two crieria for r.c.p are trivally satisfied. Hausdorff measure is the standard tool for geometric measure theory (Federer, 2014), defined as $\mathcal{H}^s(A) = \lim_{\delta \to 0} \inf\{\sum [\text{diam}(S_i)]^s : A \subseteq \cup S_i, \text{diam}(S_i) \leq \delta, \text{diam}(S_i) = \sup_{x,y \in S} \|x - y\|\}$. We denote the normalized Hausdorff measure as $\bar{\mathcal{H}}^s(A) = \frac{\Gamma(\frac{1}{2})^s}{2^s \Gamma(\frac{s}{2}+1)} \mathcal{H}^s(A)$. When s is an integer, Lebesgue and normalized Hausdorff measures coincide $\mu(A) = \bar{\mathcal{H}}^s(A)$ (Evans and Gariepy, 2015).

Similar to the proof of (2) of Proposition 2 of Diaconis et al. (2013), using co-area formula (Federer, 2014):

$$\begin{aligned} \mu(E \cap v^{-1}(F)) &= \int_{\mathbb{R}^p} \mathbb{1}_{\theta \in E} \mathbb{1}_{\theta \in v^{-1}(F)} \pi_{\mathcal{R}}(\theta) d\theta \\ &= \int_{\mathbb{R}^{p-s}} \left[\int_{v^{-1}(y)} \mathbb{1}_{\theta \in E} \mathbb{1}_{v(\theta) \in F} \frac{\pi_{\mathcal{R}}(\theta)}{J(v(\theta))} \bar{\mathcal{H}}^s(d\theta) \right] dy \\ &= \int_{F \cap \mathbb{R}^{p-s}} \left[\int_{E \cap \mathbb{R}^s} \frac{\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=y}}{J(v(\theta))} d\theta \right] dy \end{aligned} \tag{12}$$

For $y \in \{y' : m(y') = 0\}$, $\int_{E\cap\mathbb{R}^s} \frac{\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=y}}{J(v(\theta))}d\theta \leq \int_{\mathbb{R}^s} \frac{\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=y}}{J(v(\theta))}d\theta = 0$; for $y \in \{y' : m(y') = \infty\}$, since $\mu(\mathbb{R}^p) = \int \mathbb{1}_{m(y)=\infty} m(y)dy + \int \mathbb{1}_{m(y)<\infty} m(y)dy = 1$, one must have $\int \mathbb{1}_{m(y)=\infty} dy = 0$. Combining parts yields

$$\begin{aligned}
\mu(E \cap v^{-1}(F)) &= \int_F \mathbb{1}_{m(y)\in(0,\infty)} \left[\int_E \frac{\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=y}}{J(v(\theta))}d\theta\right] dy \\
&= \int_F \mathbb{1}_{m(y)\in(0,\infty)} \left[\int_E \frac{1}{m(y)}\frac{\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=y}}{J(v(\theta))}d\theta\right] m(y)dy \qquad (13) \\
&= \int_F f(y, E)\mu_y(dy)
\end{aligned}$$

□

The above remark gives the definition of r.c.p given all $v(\theta) \in Y$. Since our primary interest is when $v(\theta) = \mathbf{0}$, as long as $m_s(\mathbf{0}) \in (0, \infty)$ at certain integer s, we would have a valid conditional density

$$\pi_{\mathcal{D}}(\theta) = \frac{1}{m(\mathbf{0})}\frac{\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=\mathbf{0}}}{J(v(\theta))} \qquad (14)$$

The dimension s is often refered as 'intrinsic' dimension of $\mathcal{D}$. Formallly, one would use a standard concept in geometric measure theory named Hausdorff measure, (Federer, 2014). The d-dimensional Hausdorff measure is the limit total volume of the d-dimenional balls covering A, $\mathcal{H}^d(A) = \lim_{\delta\to 0} \inf\{\sum [\text{diam}(S_i)]^d : A \subseteq \cup S_i, \text{diam}(S_i) \leq \delta, \text{diam}(S_i) = \sup_{x,y\in S} \|x - y\|\}$. Then the intrinsic dimension is equal to Hausdorff dimension $s = \inf_{d\geq 0}\{H^d(\mathcal{D}) = 0\} = \sup_{d\geq 0}\{H^d(\mathcal{D}) = \infty\}$, at which the Hausdorff measure transitions from 0 to ∞. Although finding s can be challenging (Mardia, 1975; Bowen, 1979), one could could heuristically test $s \in \{p, p-1, \ldots, 0\}$ if s is known to be an integer. Fortunately, for posterior estimation, there is no need for estimating s or the normalizing constant $m_s(\mathbf{0})$ in Monte Carlo sampling.

We now quantify the approximation error of the exponential approximation (6) in constraint relaxing. We use $\Pi(.)$ and $\tilde{\Pi}(.)$ to represent the measures under exact and approximating distributions. For easier notation, we re-parameterize the approximating part as $\exp(-\lambda^{-1}v(\theta))$ where $\lambda = \max_k \lambda_k$ and $v(\theta) = \sum_{k=1}^d \frac{|v_k(\theta)|^\alpha}{\lambda_k^*}$ with $\lambda_k^* = \lambda_k/\lambda$ and define a conditional expectation, $\mathbb{E}(g(\theta) \mid v(\theta) = x) = \int_{\mathbb{R}^s} \frac{g(\theta)\mathbb{1}_{v(\theta)=x}\pi_{\mathcal{R}}(\theta)}{Jv(\theta)}d\theta$

We now assess the behavior of approximation error in terms of 1-Wasserstein distance, as λ decreases towards 0. The 1-Wasserstein distance $W_1(\Pi, \tilde{\Pi})$ represents the minimal amount of transport needed to transform one distribution to another. Formally, it is defined as

$$W_1(\Pi, \tilde{\Pi}) = \inf_{\gamma \in \Gamma(\Pi, \tilde{\Pi})} \int \|\theta - \theta^*\| d\gamma(\theta, \theta^*)$$

where $\Gamma(\Pi, \tilde{\Pi})$ is the family of all joint measures of with Π and $\tilde{\Pi}$ as the marginals.

Remark 2. *The 1-Wasserstein distance between the measures based on* (5) *and* (6) *has*

$$\lim_{\lambda \to 0} W_1(\Pi, \tilde{\Pi}) = 0.$$

Further, for $\alpha = 1$ *in* (6),

$$W_1(\Pi, \tilde{\Pi}) \leq \lambda(\frac{k_1 k_2}{m(0)^2} + \frac{k_1}{m(0)}) + \exp(-\lambda^{-1} t)(\frac{k_1}{m(0)^2} + \frac{k_3}{m(0)}), \tag{15}$$

where $k_1 = \sup_{g:\|g\|_L \leq 1} \sup_{t^* \in [0,t)} \|\mathbb{E}(g(\theta^*) \mid v(\theta^*) = t^*)\|$, $k_2 = \sup_{t^* \in (0,t)} m(t^*)$ *and* $k_3 = \sup_{g:\|g\|_L \leq 1} \mathbb{E}(\|g(\theta)\|)$.

Proof. Let $g : \mathbb{R}^p \to \mathbb{R}$ be a 1-Lipschitz continuous function, i.e. $\|g(x) - g(y)\| \leq \|x - y\|$, denoted by $\|g\|_L \leq 1$. By Kantorovich-Rubinstein duality, the 1-Wasserstein distance based on Euclidean metric equals to:

$$W_1(\Pi, \tilde{\Pi}) = \sup_{g:\|g\|_L \leq 1} \int g(x) \Pi(dx) - \int g(y) \tilde{\Pi}(dy) \tag{16}$$

Taking $g(\theta) = \exp(-\lambda^{-1} v(\theta))$ yields

$$\begin{aligned} m_\lambda &= \int_{\mathbb{R}} \left[\int_{v^{-1}(x)} \frac{\exp(-\lambda^{-1} v(\theta)) \pi_{\mathcal{R}}(\theta)}{Jv(\theta)} \mathcal{H}^{p-d}(d\theta) \right] \mathbb{1}_{x \geq 0} dx \\ &= \int_{\mathbb{R}} m(x) \exp(-\lambda^{-1} x) \mathbb{1}_{x \geq 0} dx. \end{aligned} \tag{17}$$

Taking $g(\theta) = \mathbb{1}_{v(\theta)=0}$ yields

$$m_0 = \int_{\mathbb{R}} \left[\int_{v^{-1}(y)} \frac{\pi_{\mathcal{R}}(\theta)}{Jv(\theta)} \mathcal{H}^{p-d}(d\theta) \right] \mathbb{1}_{y=0} dy = \int_{v^{-1}(0)} \frac{\pi_{\mathcal{R}}(\theta)}{Jv(\theta)} \mathcal{H}^{p-d}(d\theta) = m(0) \tag{18}$$

Clearly $m_\lambda \geq m_0$.

1. Asymptotic result:

We have

$$
\begin{aligned}
&\sup_{g:\|g\|_L\le 1}\int g(\theta)\left[\frac{\exp(-\lambda^{-1}v(\theta))}{m_\lambda}-\frac{\mathbb{1}_{v(\theta)=0}}{m_0}\right]\frac{\pi_{\mathcal{R}}(\theta)}{Jv(\theta)}\mathcal{H}^{p-d}(d\theta)\\
&=\sup_{g:\|g\|_L\le 1}\int_{\mathbb{R}}\mathbb{E}(g(\theta)\mid x)\left[\frac{\exp(-\lambda^{-1}x)\mathbb{1}_{x\ge 0}}{m_\lambda}-\frac{\mathbb{1}_{x=0}}{m_0}\right]dx\\
&=\sup_{g:\|g\|_L\le 1}\int_{\mathbb{R}}\mathbb{E}(g(\theta)\mid x)\left[\frac{1}{m_\lambda}-\frac{1}{m_0}\right]\mathbb{1}_{x=0}dx+\sup_{g:\|g\|_L\le 1}\int_{\mathbb{R}}\mathbb{E}(g(\theta)\mid x)\frac{\exp(-\lambda^{-1}x)}{m_\lambda}\mathbb{1}_{x>0}dx\\
&\le\sup_{g:\|g\|_L\le 1}\|\mathbb{E}(g(\theta)\mid 0)\|\left[\frac{1}{m_0}-\frac{1}{m_\lambda}\right]+\frac{1}{m_0}\sup_{g:\|g\|_L\le 1}\int_{\mathbb{R}}\|\mathbb{E}(g(\theta)\mid x)\|\exp(-\lambda^{-1}x)\mathbb{1}_{x>0}dx
\end{aligned}
\tag{19}
$$

Note $m_\lambda\le\int_{\mathbb{R}}m(x)\mathbb{1}_{x\ge 0}dx=\int_{\mathbb{R}}\pi_{\mathcal{R}}(\theta)=1$. By dominated convergence theorem,

$$
\lim_{\lambda\to 0}m_\lambda=\int_{\mathbb{R}}m(x)\lim_{\lambda\to 0}\exp(-\lambda^{-1}x)\mathbb{1}_{x\ge 0}dx=m_0. \tag{20}
$$

Since $\sup_{g:\|g\|_L\le 1}\int_{\mathbb{R}}\|\mathbb{E}(g(\theta)\mid x)\|\exp(-\lambda^{-1}x)dx\le\int_{\mathbb{R}}\sup_{g:\|g\|_L\le 1}\|\mathbb{E}(g(\theta)\mid x)\|\exp(-\lambda^{-1}x)dx$, letting $q_\lambda=\sup_{g:\|g\|_L\le 1}\|\mathbb{E}(g(\theta)\mid x)\|\exp(-\lambda^{-1}x)\mathbb{1}_{x>0}$, we have $0\le q_1-q_{\lambda_1}\le q_1-q_{\lambda_2}$ for $1\ge\lambda_1\ge\lambda_2$, by monotone convergence theorem, $\lim_{\lambda\to 0}\int[q_1(x)-q_\lambda(x)]dx=\int[q_1(x)-q_0(x)]dx$ hence $\lim_{\lambda\to 0}\int q_\lambda(x)dx=0$. Combining the results yields

$$
\lim_{\lambda\to 0}W_1(\Pi,\tilde{\Pi})=0. \tag{21}
$$

2. Non-asympotic result:

$$
\begin{aligned}
\frac{1}{m_0}-\frac{1}{m_\lambda}&\le\frac{\int_{\mathbb{R}}m(x)\exp(-\lambda^{-1}x)\mathbb{1}_{x>0}dx}{m_0^2}\\
&=\frac{1}{m_0^2}\left[\int_0^t m(x)\exp(-\lambda^{-1}x)dx+\int_t^\infty m(x)\exp(-\lambda^{-1}x)dx\right]\\
&\le\frac{1}{m_0^2}\left[\sup_{t^*\in(0,t)}m(t^*)\int_0^t\exp(-\lambda^{-1}x)dx+\exp(-\lambda^{-1}t)\int_t^\infty m(x)dx\right]\\
&\le\frac{1}{m_0^2}\left[\lambda\sup_{t^*\in(0,t)}m(t^*)+\exp(-\lambda^{-1}t)\right]
\end{aligned}
\tag{22}
$$

$$
\begin{aligned}
&\sup_{g:\|g\|_L\leq 1}\int_{\mathbb{R}}\|\mathbb{E}(g(\theta)\mid x)\|\exp(-\lambda^{-1}x)\mathbb{1}_{x>0}dx\\
&\leq \sup_{g:\|g\|_L\leq 1}\sup_{t^*\in(0,t)}\|\mathbb{E}(g(\theta)\mid t^*)\|\int_0^t\exp(-\lambda^{-1}x)dx+\exp(-\lambda^{-1}t)\sup_{g:\|g\|_L\leq 1}\int_t^\infty\|\mathbb{E}(g(\theta)\mid x)\|dx \qquad (23)\\
&\leq \sup_{g:\|g\|_L\leq 1}\sup_{t^*\in(0,t)}\|\mathbb{E}(g(\theta)\mid t^*)\|\lambda+\exp(-\lambda^{-1}t)\sup_{g:\|g\|_L\leq 1}\mathbb{E}(\|g(\theta)\|)
\end{aligned}
$$

Combining (19)(22)(23), $k_1=\sup_{g:\|g\|_L\leq 1}\sup_{t^*\in[0,t)}\|\mathbb{E}(g(\theta)\mid t^*)\|$, $k_2=\sup_{g:\|g\|_L\leq 1}\mathbb{E}(\|g(\theta)\|)$, $k_3=\sup_{t^*\in(0,t)}m(t^*)$

$$
\begin{aligned}
&\sup_{g:\|g\|_L\leq 1}\int g(x)\Pi(dx)-\int g(x)\tilde{\Pi}(dx)\\
&\leq \lambda(\frac{k_1k_3}{m_0^2}+\frac{k_1}{m_0})+\exp(-\lambda^{-1}t)(\frac{k_1}{m_0^2}+\frac{k_2}{m_0})
\end{aligned}
\qquad (24)
$$

□

The first part shows the asymptotic accuracy of the approximation. The second part shows the rate with non-asymptotic λ under mild assumptions. The interpretation for these assumptions is that if in a small space expansion of $\mathcal{D}$, defined as $\{\theta^* : v(\theta^*)\in[0,t]\}$, the marginal density of $v(\theta^*)$ and the conditional expectation of Lipschitz functions are bounded $k_1,k_2=\mathcal{O}(1)$, and the expected norm of Lipschitz function are smaller than a bound that grows near exponentially $k_3=\mathcal{O}(\lambda\exp(t/\lambda))$, then the distance $W_1(\Pi,\tilde{\Pi})$ converges to 0 in $\mathcal{O}(\lambda)$ as $\lambda\to 0$.

3 Posterior Computation

Conditioning the unconstrained density onto $\mathcal{D}$ often disrupts the posterior conjugacy. Thanks to the constraint relaxation via approximation or reparameterization, we can sample the posterior directly inside $\mathcal{R}$. One can exploit conventional sampling tools such as slice sampling, adaptive Metropolis-Hastings and Hamiltonian Monte Carlo (HMC). In this section, we focus on HMC for its easiness to use and good performance in block sampling with relatively high dimension.

3.1 Hamiltonian Monte Carlo under Constraint Relaxation

We provide a brief overview of HMC for continuous θ^* under constraint relaxation. Discrete extension is possible via recent work of Nishimura et al. (2017). For easy notation, we use q to represent θ^* under approximation (6), and $\{\theta^*, w\}$ under reparameterization (7).

In order to sample q, HMC introduces an auxillary momentum variable $p \sim \text{No}(0, M)$. The covariance matrix M is referred to as a *mass matrix* and is typically chosen to be the identity or adapted to approximate the inverse covariance of q. HMC then sample from the joint target density $\pi(q,p) = \pi(q)\pi(p) \propto \exp(-H(q,p))$ where, in the case of the posterior under relaxation,

$$\begin{aligned} &H(q,p) = U(q) + K(p), \\ &\text{where } U(q) = -\log \pi(q), \\ &\qquad K(p) = \frac{p' M^{-1} p}{2}. \end{aligned} \tag{25}$$

with $\pi(q)$ is the unnormalized density in (6) or (7).

From the current state $(q^{(0)}, p^{(0)})$, HMC generates a proposal for Metropolis-Hastings algorithm by simulating Hamiltonian dynamics, which is defined by a differential equation:

$$\begin{aligned} \frac{\partial q^{(t)}}{\partial t} &= \frac{\partial H(q,p)}{\partial p} = M^{-1} p, \\ \frac{\partial p^{(t)}}{\partial t} &= -\frac{\partial H(q,p)}{\partial q} = -\frac{\partial U(q)}{\partial q}. \end{aligned} \tag{26}$$

The exact solution to (26) is typically intractable but a valid Metropolis proposal can be generated by numerically approximating (26) with a reversible and volume-preserving integrator (Neal, 2011). The standard choice is the *leapfrog* integrator which approximates the evolution $(q^{(t)}, p^{(t)}) \to (q^{(t+\epsilon)}, p^{(t+\epsilon)})$ through the following update equations:

$$p \leftarrow p - \frac{\epsilon}{2}\frac{\partial U}{\partial q}, \quad q \leftarrow q + \epsilon M^{-1} p, \quad p \leftarrow p - \frac{\epsilon}{2}\frac{\partial U}{\partial q} \tag{27}$$

Taking L leapfrog steps from the current state $(q^{(0)}, p^{(0)})$ generates a proposal $(q^*, p^*) \approx (q^{(L\epsilon)}, p^{(L\epsilon)})$, which is accepted with the probability

$$1 \wedge \exp\Big(-H(q^*, p^*) + H(q^{(0)}, p^{(0)}))\Big)$$

3.2 Computing Efficiency and Support Expansion

Since constraint relaxation expands the support from $\mathcal{D}$ to $\mathcal{R}$, it is useful to study the effect of space expansion on the computing efficiency. In this section, we provide some heuristic quantification of the effects and provide a practical guidance on choosing $\pi(w)$ or λ in the two constraint relaxation strategies.

In understanding the computational efficiency of HMC, it is useful to consider the number of leapfrog steps to be a function of ϵ and set $L = \lfloor \tau/\epsilon \rfloor$ for a fixed integration time $\tau > 0$. In this case, the mixing rate of HMC is completely determined by τ in the limit $\epsilon \to 0$ (Betancourt, 2017). In practice, while a smaller stepsize ϵ leads to a more accurate numerical approximation of Hamiltonian dynamics and hence a higher acceptance rate, it takes a larger number of leapfrog steps and gradient evaluations to achieve good mixing. For an optimal computational efficiency of HMC, therefore, the stepsize ϵ should be chosen only as small as needed to achieve a reasonable acceptance rate (Beskos et al., 2013; Betancourt et al., 2014). A critical factor in determining a reasonable stepsize is the *stability limit* of the leapfrog integrator (Neal, 2011). When ϵ exceeds this limit, the approximation becomes unstable and the acceptance rate drops dramatically. Below the stability limit, the acceptance rate $a(\epsilon)$ of HMC increases to 1 quite rapidly as $\epsilon \to 0$ and in fact satisfies $a(\epsilon) = 1 - \mathcal{O}(\epsilon^4)$ (Beskos et al., 2013).

For simplicity, the following discussions assume the mass matrix M is taken to be the identity. Let $\mathbf{H}_U(q)$ denote the hessian matrix of $U(q) = -\log \pi(q)$ and let $\xi_1(q)$ denotes the first largest eigenvalue of $\mathbf{H}_U(q)$. While analyzing stability and accuracy of an integrator is highly problem specific, the linear stability analysis and empirical evidences suggest that, for stable approximation of Hamiltonian dynamics by the leapfrog integrator in $\mathbb{R}^p$, the condition $\epsilon < 2\xi_1(\theta)^{-1/2}$ must hold on most regions of the parameter space (Hairer et al., 2006). Besides the eigenvalue, if the support of q is a constrained space $\mathcal{Q}$, another limiting factor is roughly the shortest distance to the boundary $\eta(\theta; \mathcal{Q}) = \inf_{q' \notin \mathcal{Q}} \|q' - q\|$. If either $\eta(\theta; \mathcal{Q})$ or $\xi_1(\theta)^{-1/2}$ is close to 0, the upper bound would be too low to obtain efficient computation. In constrained model, the parameter space $\mathcal{D}$ can have very small $\eta(\theta; \mathcal{D})$. Constraint relaxation can reduce this problem as it provides support expansion.

For approximation (6), to minimize approximation error, one can choose to relax a subset of constraints. Observing $\mathcal{D} = \cap D_k$, each approximation $\exp(-\frac{|v_k(\theta^*)|^\alpha}{\lambda_k})$ corresponds to a constrained space $\mathcal{D}_k$. The practical strategy is that, for $\mathcal{D}_k$'s with $\eta(\theta; \mathcal{D}_k) \approx 0$, one use moderate λ_k to induce some support expansion (denoted by $\lambda_k \geq \zeta$ with ζ moderately small but not very close to 0); for $\mathcal{D}_k$'s without this issue, one uses very small $\lambda_k \approx 0$ to almost always uphold the constraint. The latter was also suggested by Neal (2011) as creating a high 'energy wall'. To reduce the inaccuracy of HMC near the boundary with $\lambda_k \approx 0$, we use random step size ϵ at each iteration.

The Hessian $\mathbf{H}_U(q)$ under this strategy is given by

$$\mathbf{H}_U(q) = -\mathbf{H}_{\log L(y;\theta^*)\pi_{\mathcal{R}}(\theta^*)/J(v(\theta^*))}(\theta^*) + \sum_k \lambda_k^{-1} \mathbf{H}_{|v_k|^\alpha}(\theta^*) \mathbb{1}_{\theta \notin \mathcal{D}_k}, \tag{28}$$

where the second term $\lambda_k^{-1}\mathbf{H}_{v_k}(\theta)\mathbb{1}_{\theta \notin \mathcal{D}_k}$ is 0 unless $\theta \notin \mathcal{D}_k$. Since λ_k^{-1}'s in the second term often dominate

the eigenvalue, hence $\xi_1^{-1/2}(\theta^*) \approx \min_{\lambda_k \geq \zeta} \lambda_k^{1/2}$. A trade-off between approximation accuracy and computational efficiency is involved. Fortunately, the approximation error $\mathcal{O}(\max_{\lambda_k \geq \zeta} \lambda_k)$ decreases faster than the efficiceny cap $\mathcal{O}(\min_{\lambda_k \geq \zeta} \lambda_k^{1/2})$. In our experiment, we did find changing from $\lambda_k = 10^{-4}$ to 10^{-5} requires approximately 3 times of computing budget, due the effect on stability bound.

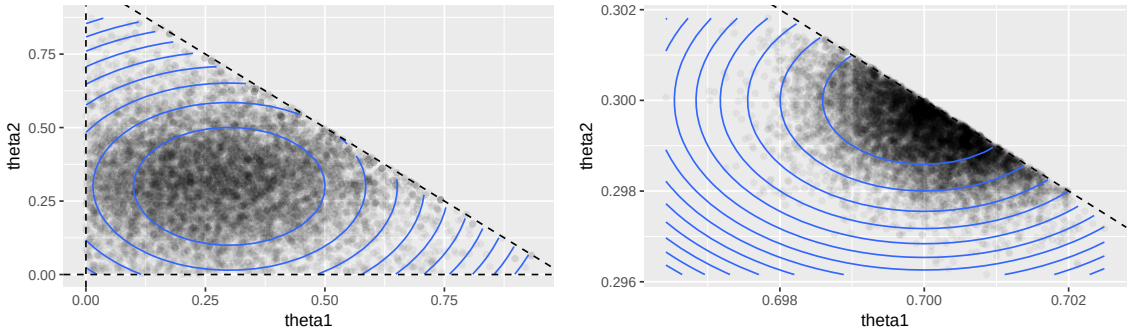
On the other hand, since reparameterization (7) does not involve such trade-off, it is more preferrable despite its less general application. Letting $\mathcal{Q}_\theta \subset \mathcal{D}$ denote the support for the constrained $\theta \in \mathcal{D}$, the reparameterization changes the support to $Q_{\theta^*} = \{g^{-1}(\theta; w) : \theta \in \mathcal{Q}\}$. Therefore, one could choose $\pi(w)$ to substantially increase $\eta(\theta^*; Q_{\theta^*})$, provided it does not significantly reduce $\xi_1^{-1/2}(\theta^*)$. In literature related to parameter expansion Gibbs sampling (Liu and Wu, 1999), it was suggested to use flat improper prior for $\pi(w)$; instead, since HMC is used, we suggest assigning $\pi(w)$ loosely centered at $\mathbf{0}$ (corresponding to $\theta^* = \theta$), which makes θ^* a mild relaxation of θ and ensures no substantial change in $\xi_1^{-1/2}(\theta^*)$. More will be illustrated in the next section.

4 Simulations

In this section, we run simulations on several toy examples, mainly to compare with existing approaches on computing efficiency and provide empirical evidence supporting our previous result.

4.1 Gaussian under Linear Inequality

We first consider linear models under linear inequality constraints. Although recent work proposed a new class of prior with posterior conjugacy (Danaher et al., 2012) , one can simply use consider Gaussian prior under our framework. To illustrate the computing performance, we sample a bivariate Gaussian $\theta \sim \text{No}\left(\mu, I\sigma^2\right)$ subject to linear inequality $\theta \in (0,1)^2, \theta_1 + \theta_2 < 1$, which is a triangle. In two separate settings, choosing (μ, σ^2) as $([0.3, 0.3], 1/10)$ induces a wide-spread distribution centered in the interior of $\mathcal{D}$, while $([0.7, 0.3]', 1/10^4)$ induces a distribution concentrated on the boundary of $\mathcal{D}$. For constraint relaxation, we use the approximation form (6) with $\exp(-\frac{|v(\theta)|}{\lambda})$, with $v(\theta) = |\theta_1 + \theta_2 - 1|_+ + |-\theta_1|_+ + |-\theta_2|_+ + |\theta_1 - 1|_+ + |\theta_2 - 1|_+$. Since the triangle has $\eta(\theta; \mathcal{D})$ away from 0, small $\lambda = 10^{-8}$ allows almost no approximation error. Figure 2 plots the posterior sample and its contour. To compare, we ran simple rejection sampling with untruncated normal proposal $\text{No}(\mu, I\sigma^2)$. It suffers from a rapidly growing rejection rate from 12% to 51%, as μ moves further away from the center of $\mathcal{D}$.



(a) Constrained $\text{No}([0.3, 0.3], 1/10)$ (b) Constrained $\text{No}([0.7, 0.3], 1/10^4)$

Figure 2: Posterior sample of bivariate normal distribution subject to linear inequality constraints $\theta \in (0,1)^2, \theta_1 + \theta_2 < 1$, using HMC with constraint relaxation. Posterior is spread out around the center (panel (a)) or concentrated on the boundary (panel (b)) of the region.

4.2 von Mises–Fisher on Unit Circle

To illustrate equality constraint relxation, we generate a simple von Mises–Fisher distribution $\pi_{\mathcal{D}}(\theta) \propto \exp(F'\theta)$ on a unit circle $\{(\theta_1, \theta_2) : \theta_1^2 + \theta_2^2 = 1\}$. We use $F = (5, 5)$ to induce a relatively spread-out θ on the manifold. For sampling, we compare three strategies: approximate constraint relaxation using $\exp(-\frac{|\theta'\theta - 1|}{\lambda})$ for the indicator, reparameterization constraint relaxation using $\theta_1 = \frac{\theta_1^*}{w+1}, w + 1 = \sqrt{(\theta_1^*)^2 + (\theta_2^*)^2}$ and $\pi(w) \sim \text{No}(0, 1)$ and exact von Mises–Fisher obtained using 'movMF' package.

Unlike the previous linear inequality constraint, the unit circle has $\eta(\theta; \mathcal{D}) = 0$ for all $\theta \in \mathcal{D}$, therefore, some support expansion is needed for HMC. We test $\lambda = 10^{-3}$, 10^{-4} and 10^{-5} for approximation constraint relaxation. To compare the efficiency of HMC, we fix the number of leap-frog steps to 20 within one iteration HMC, and let software STAN automatically tune for stable step size. Table 1 shows the effective sample size per 1000 iterations, the effective 'violation' $|v(\theta)| = |\theta_1 + \theta_2 - 1|$ and the 1-Wasserstein distance W_1 as the approximation error. As W_1 is numerically computed, to provide a baseline error, we also calculate the average W_1 comparing two independent samples from the same exact distribution. The approximation error W_1 based on $\lambda = 10^{-5}$ approximation is indistinguishable from this low numerical error, while the other approximations have slightly larger error but more effective samples.

	Constraint Relaxation HMC				Exact
	Approximation			Reparameterization	
	$\lambda = 10^{-3}$	$\lambda = 10^{-4}$	$\lambda = 10^{-5}$		
W_1	0.050 (0.019, 0.095)	0.034 (0.027, 0.037)	0.014 (0.013,0.025)	0.017 (0.0012,0.026)	0.015 (0.0014,0.025)
$\lvert v(\theta)\rvert \mid y$	9×10^{-4} $(2.6 \cdot 10^{-5}, 3.3 \cdot 10^{-3})$	9×10^{-5} $(2.0 \cdot 10^{-6}, 3.4 \cdot 10^{-4})$	9×10^{-6} $(2.7 \cdot 10^{-7}, 3.5 \cdot 10^{-5})$	0	0
ESS /1000 Iterations	751.48	260.54	57.10	788.30	

Table 1: Benchmark of constraint relaxation methods on sampling von–Mises Fisher distribution on a unit circle. For each approximation, average approximation error (with 95% credible interval, out of 10 repeated experiments) is computed, and numeric error of W_1 is shown under column 'exact' as comparing two independent copies from the exact distribution.. Effective sample size shows reparameterization and approximation with relatively large λ have high computing efficiency.

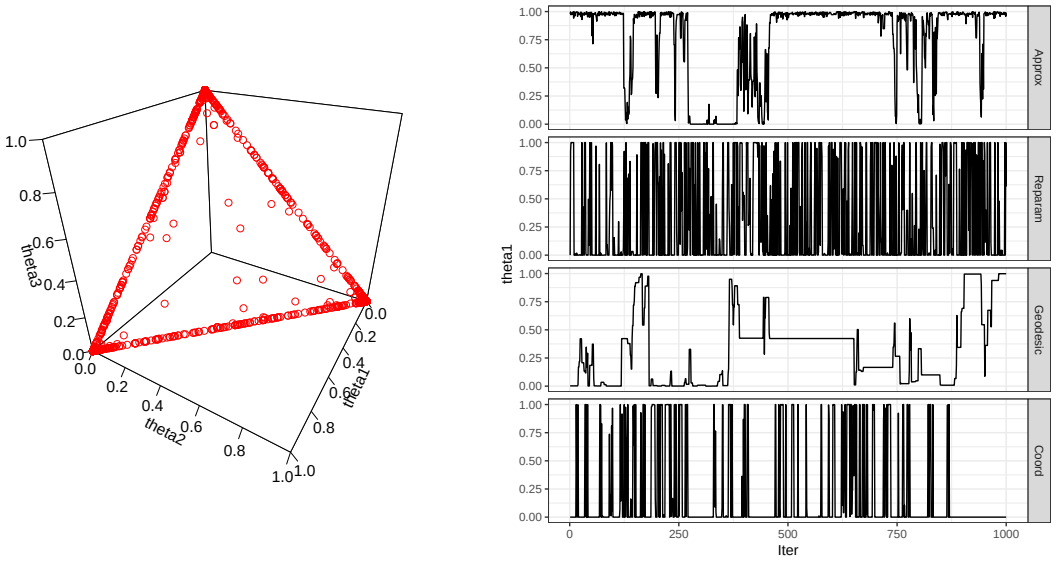
4.3 Dirichlet on a Simplex

Lastly, we experiment with a particularly challenging distribution on a $(p-1)$-simplex, defined by $\{\theta : \theta \in (0,1)^p, \sum_{i=1}^p \theta_i = 1\}$. We consider Dirichlet distribution $\text{Dir}(\alpha)$, with $\pi_{\mathcal{D}}(\theta) \propto \prod_{i=1}^p \theta_i^{\alpha-1}$. When the concentration parameter $\alpha < 1$, $\text{Dir}(\alpha)$ exhibits sparse property that some θ_i's become very close to 0, which is exploited in many modeling approaches CITES. Despite the simple form, the computation can be quite difficult if there is large uncerntainty associated with θ on top of sparsity. The distribution will be multi-modal with distribution scattered along the boundary of the simplex (Figure 3(a)).

To illustrate, we consider $p = 3$ and various values of $\alpha \in \{1, 0.5, 0.1, 0.01\}$. We test the performance of constraint relaxation under approximation and reparameterization. To compare, we also test the standard HMC using coordinate system $\theta_1 = \cos^2(\theta_1^*), \theta_2 = \sin^2(\theta_1^*)\cos^2(\theta_2^*), \theta_3 = \sin^2(\theta_1^*)\sin^2(\theta_2^*)$ for $\theta^* \in [0, 2\pi)^2$, which is equivalent to stick-breaking representation CITES; and the geodesic HMC utilizing the geometric flow directly on the simplex (Byrne and Girolami, 2013). For all HMCs, we fix the number of steps in each iteration to be 30 and tune the step size.

Table 2 lists the effective sample sizes under different α's. As α becomes smaller than 1, approximation HMC based on constraint relaxation and geodesic HMC become worse in performance, while reparameterizations based on constraint relaxation or coordinate system are much less impacted. Figure 3(b) shows at $\alpha = 0.01$, the approximate HMC and geodesic HMC are stuck for a long time, while reparameterizations based on constraint relaxation works substantially better. As a well-tested reparameterization, it is not surprising that HMC based on coordinate system still works acceptably well in this case.

The difficulty that approximate HMC based on constraint relxation encountered is well anticipated. Byrne and Girolami (2013) have previously reported similar slow-down of geodesic HMC computing on hyper-Dirichlet distribution CITES with $\alpha < 1$. And our approximate HMC can also be viewed as an approximation to the exact geodesic HMC, with the metric tensor product replaced by our exponential function.



(a) 2,000 samples from Dir(0.01) on 2-simplex.

(b) Traceplot of θ_1 using 4 types of HMCs.

Figure 3: Sampling of Dirichlet on an simplex with distribution concentrated on the boundaries. Panel(a) illustrates the distribution under Dir(0.01); Panel(b) compares the traceplots of 4 different types of HMCs, which are based on: approximation using constraint relaxation with $\lambda = 10^{-3}$, reparameterization using constraint relaxation, geodesic flow on simplex (Byrne and Girolami, 2013) and coordinate system.

	Constraint Relaxation HMC			Geodesic HMC	Coordinate HMC
	Approx $\lambda = 10^{-3}$	Approx $\lambda = 10^{-4}$	Reparameterization		
ESS /1000 Iter. ($\alpha = 1$)	511.43	146.07	947.53	174.14	**961.08**
ESS /1000 Iter. ($\alpha = 0.5$)	145.15	33.16	**912.94**	31.47	846.92
ESS /1000 Iter. ($\alpha = 0.1$)	88.32	26.88	**992.75**	28.70	875.83
ESS /1000 Iter. ($\alpha = 0.01$)	20.54	3.91	**722.44**	17.26	128.55

Table 2: Average effective sample size per 1000 iterations in Dir(α), under different α.

5 Application: Sparse Basis Learning in Network Analysis

We now consider a real data application in brain network analysis. The brain connectivity structures are obtained in the data set KKI-42 (Landman et al. 2011), which consists of $n = 21$ healthy subjects without any history of neurological disease. We take the first scan out of the scan-rescan data as the input. Each observation is a $V \times V$ symmetric network A_i, recorded as adjacency matrix A_i for $i = 1, \ldots, n$. The regions are constructed via the Desikan et al. (2006) atlas, for a total of V = 68 nodes. For the ith matrix A_i, $A_{i,k,l} \in \{0, 1\}$ is the element on the kth row and lth column of A_i, with $A_{i,k,l} = 1$ indicating there is an connection betwen kth and lth region, $A_{i,k,l} = 0$ if there is no connection. The matrix is symmetric due to the undirectedness of the network, but the diagonal records $A_{i,k,k}$ for all i and k are missing due to the lack of meaning in self-connectivity. For this reason, in addition to the sample size n much smaller than the ambient space $p = \frac{V(V-1)}{2}$, a probabilistic Bayesian model can be parituclarly useful for uncertainty quantification.

One scientific interest in neuroscience is to identify the important regions (nodes) in the formation of the

average brain network. Although probabilistic principal component analysis (PPCA) are routinely used for extracting the first few important bases, it does not selectively identify the important nodes. To achieve this goal, one natural way is to apply shrinkage on the elements of principal components.

Geometrically, the principle components denoted by $\{U_1, \ldots, U_d\}$ reside on a Stiefel manifold $\mathcal{V}(d, n) = \{U : U'U = I_d\}$, where $U = [U_1, \ldots, U_d]$ is the $n \times d$ matrix. Using r to index $1, \ldots, d$, each U_r represents a $(n-1)$-hypersphere. Applying shrinkage forces some of its sub-coordinates to be close to 0, which is approximately reducing U_r to a lower-dimensional hyper-sphere. Although there was previous work using sparse PCA (Zou et al., 2006) on continuous outcome, little work has been done on a probabilistic model for binary outcome.

For shrinkage, we utilize the double Pareto prior (Armagan et al., 2013) as unconstrained distribution $\pi_{\mathcal{R}}$ and adapt it onto Stiefel manifold via (2). For simplicity, we consider a probabilistic PCA model with a random intercept ,

$$A_{i,k,l} \sim \text{Bern}(\frac{1}{1+\exp(-\psi_{k,l} - Z_i)})$$

$$\psi_{k,l} = \sum_{r=1}^{d} v_r u_{k,r} u_{l,r}$$

$$U'U = I \text{ with } U = \{u_{k,r}\}_{k=1,\ldots,n;r=1,\ldots,d}$$

$$\pi(u_{k,r}) \propto \frac{1}{2(1+|u_{k,r}|)^2}$$

$$Z_i \sim \text{No}(0, \sigma_z^2), \quad \sigma_z^2 \sim \text{IG}(2,1)$$

$$v_r \sim \text{No}(0, \sigma_v^2), \quad \sigma_v^2 \sim \text{IG}(2,1)$$

for $k > l$, $k = 2, \ldots, V$, $i = 1, \ldots, n$. We adapt $\pi_{\mathcal{R}}(u)$ as the standard double Pareto prior (Armagan et al., 2013), to avoid dealing with the normalizing constant due to constraint; Z_i is a scalar used as random intercept; for scale parameter, we choose weakly informative prior inverse Gamma $\text{IG}(2,1)$, as appropriate for the scale under the logistic link.

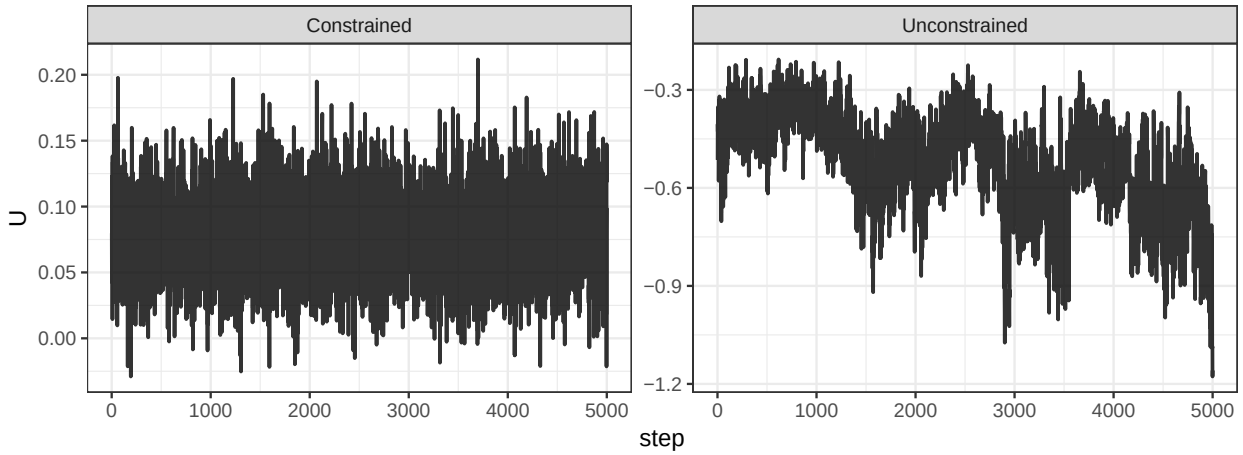
This model is a special Tucker decomposition with a sparse core tensor, whose diagonal plane is equal to D and 0 for other elements. The Tucker decomposition is more flexible than another routinely used decomposition, namely parallel factor analysis (PARAFAC). The PARAFAC assumes all ranks are equal and the core tensor D only has non-zero value when all its sub-indices are equal. In this case, PARAFAC would assume $d_1 = d_2$. The additional flexibility in the Tucker is appealing, as one could utilize the varying rank over different sub-direction (mode) of the tensor. On the other hand, a completely unconstrained Tucker decomposition is not identifiable in the matrices and core tensor, due scaling. For example, one can

multiply a $d_1 \times d_1$ non-zero diagonal matrix R, to U and obtain $U^* = UR$ obtain $D^*_{\cdot,r_2,\cdot} = R^{-1}D_{\cdot,r_2,\cdot}R^{-1}$ for $r_2 = 1, \ldots, d_2$. This leaves the likelihood unchanged, creating identifiability issue.

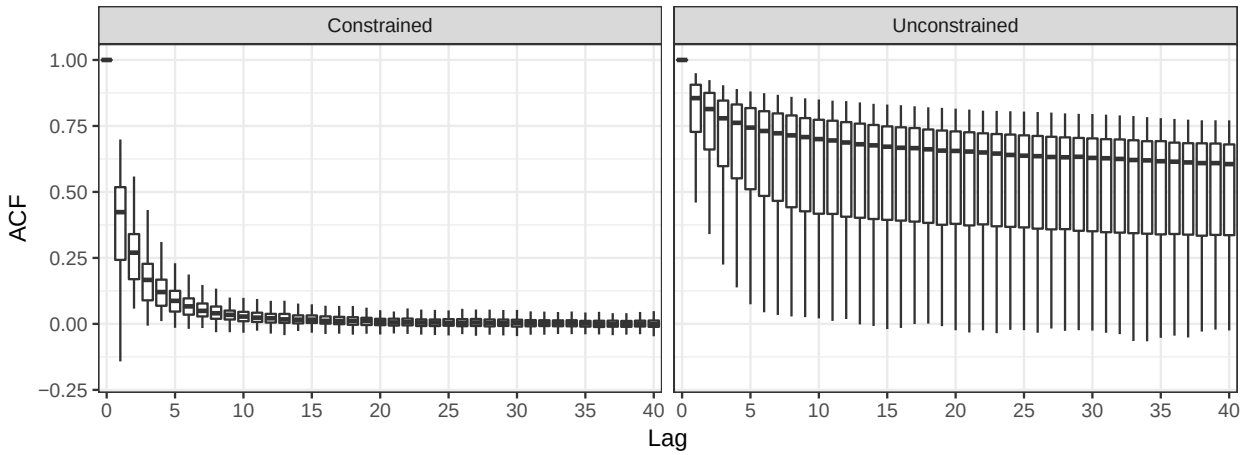
Therefore, we consider applying some constraint on the Tucker decomposition. Motivated by high-order singular value decomposition, we impose orthonormality constraints $U'U = I_{d_1}$ and $W'W = I_{d_2}$. Hoff et al. (2016) previously obtained conjugated posterior for Tucker decomposition under orthonormality constraint, however, the symmetry in undirectedness of networks breaks the conjugacy.

We assign normal prior for $U_{k,r_2} \sim \text{No}(0, \phi_1)$, $W_{i,r_1} \sim \text{No}(0, \phi_2)$, $Z_{k,l} \sim \text{No}(0, \phi_3)$, $D_{r_1,r_2} \sim No(0, \phi_{4,r_1,r_2})$ for all i, k, l, r_1, r_2, and inverse-Gamma prior $\phi_1, \phi_2, \phi_3 \overset{indep}{\sim} \text{IG}(2, 1)$, $\phi_{4,r_1,r_2} = \tau_{r_1}\tau_{r_2}$, with $\tau_{r_1}, \tau_{r_2} \overset{indep}{\sim} \text{IG}(2, 1)$ for all r_1, r_2. We

To allow estimation for model with orthonormality constraint, we use extrinsic prior with $\mathcal{K}(\theta) = \exp(-\frac{(U'U-I_{d_1})^2+(W'W-I_{d_2})^2}{\lambda})$ and set $\lambda = 10^{-3}$. To compare, we also test with the same model configuration without the orthonormality constraint. We run both models for $10,000$ steps and discard the first $5,000$ steps. Figure 4 plots the traceplot and autocorrelation for matrix U. Unconstrained base model has severe convergence issue due to the non-identifiability, while constrained model converges and show low autocorrelation for all the parameters.



(a) Traceplot of $U_{1,1}$.



(b) ACF of all elements in U

Figure 4: Orthonormality constraint in the tensor decomposition modelallows convergence and rapid mixing on the factor matrix (left column); whereas unconstrained model does not converge due to free scaling. Traceplot for one parameter in factor matrix U and boxplot for autocorrelations of all parameters are shown.

6 Discussion

The estimation difficulty associated with parameter constraint often hinders the development of new models. Often one needed to carefully avoid models without conjugate posteriors, or skillfully re-parameterize the model for a more tractable algorithm. The extrinsic approach we introduced significantly reduces the burden. Through space expansion, it allows conventional toolbox such as HMC to be easily adopted to sample posterior without closed-forms. This allows researchers to impose constraints more freely in modeling and simplifies the way to incorporate constraint information about the functional of parameters.

We show the approximation error of the extrinsic approach can controlled via tuning parameter, with some trade-off between computing time and accuracy. A potentially more efficient strategy would be obtaining a rough approximate first in $\mathcal{R}$, then projecting into $\mathcal{D}$. Lin et al. (2016) developed algorithms similar to this idea and obtained consistency result for point estimation. A useful task would be to find an optimal projection

also quantifying the uncertainty associated with finite sample. Lastly, the normalization of parameters over constrained space can sometime yield intractable integral, known as 'doubly stochastic' problem. We expect that the proposed extrinsic prior can be adapted and used together with the various existing solutions (Rao et al., 2016; Stoehr et al., 2017).

References

Armagan, A., D. B. Dunson, and J. Lee (2013). Generalized double pareto shrinkage. *Statistica Sinica 23*(1), 119.

Beskos, A., N. Pillai, G. Roberts, J. M. Sanz-Serna, and A. Stuart (2013, 11). Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli 19*(5A), 1501–1534.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434*.

Betancourt, M., S. Byrne, and M. Girolami (2014). Optimizing the integrator step size for Hamiltonian Monte Carlo. *arXiv:1411.6669*.

Bowen, R. (1979). Hausdorff dimension of quasicircles. *Publ. math. IHES 50*(1), 11–25.

Boyd, S. and L. Vandenberghe (2004). *Convex optimization*. Cambridge university press.

Byrne, S. and M. Girolami (2013). Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics 40*(4), 825–845.

Danaher, M. R., A. Roy, Z. Chen, S. L. Mumford, and E. F. Schisterman (2012). Minkowski–weyl priors for models with parameter constraints: an analysis of the biocycle study. *Journal of the American Statistical Association 107*(500), 1395–1409.

Diaconis, P., S. Holmes, M. Shahshahani, et al. (2013). Sampling from a manifold. In *Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton*, pp. 102–125. Institute of Mathematical Statistics.

Do Carmo, M. P. (2016). *Differential Geometry of Curves and Surfaces: Revised and Updated Second Edition*. Courier Dover Publications.

Dunson, D. B. and B. Neelon (2003). Bayesian inference on order-constrained parameters in generalized linear models. *Biometrics 59*(2), 286–295.

Evans, L. C. and R. F. Gariepy (2015). *Measure theory and fine properties of functions*. CRC press.

Federer, H. (2014). *Geometric measure theory*. Springer.

Gelfand, A. E., A. F. Smith, and T.-M. Lee (1992). Bayesian analysis of constrained parameter and truncated data problems using gibbs sampling. *Journal of the American Statistical Association 87*(418), 523–532.

Girolami, M. and B. Calderhead (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(2), 123–214.

Gunn, L. H. and D. B. Dunson (2005). A transformation approach for incorporating monotone or unimodal constraints. *Biostatistics 6*(3), 434–449.

Hairer, E., C. Lubich, and G. Wanner (2006). *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer-Verlag.

Hoff, P. D. (2009). Simulation of the matrix bingham–von mises–fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics 18*(2), 438–456.

Hoff, P. D. et al. (2016). Equivariant and scale-free tucker decomposition models. *Bayesian Analysis 11*(3), 627–648.

Khatri, C. and K. Mardia (1977). The von mises-fisher matrix distribution in orientation statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 95–106.

Kolmogorov, A. N. (1950). Foundations of the theory of probability.

Leao Jr, D., M. Fragoso, and P. Ruffino (2004). Regular conditional probability, disintegration of probability and radon spaces. *Proyecciones (Antofagasta) 23*(1), 15–29.

Lin, L. and D. B. Dunson (2014). Bayesian monotone regression using gaussian process projection. *Biometrika*.

Lin, L., B. St Thomas, H. Zhu, and D. B. Dunson (2016). Extrinsic local regression on manifold-valued data. *Journal of the American Statistical Association* (just-accepted).

Liu, J. S. and Y. N. Wu (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association 94*(448), 1264–1274.

Mardia, K. V. (1975). Statistics of directional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 349–393.

Nash, J. (1954). C1 isometric imbeddings. *Annals of mathematics*, 383–396.

Neal, R. M. (2011). Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo 2*, 113–162.

Nishimura, A., D. Dunson, and J. Lu (2017). Discontinuous hamiltonian monte carlo for sampling discrete parameters. *arXiv preprint arXiv:1705.08510*.

Rao, V., L. Lin, and D. B. Dunson (2016). Data augmentation for models based on rejection sampling. *Biometrika 103*(2), 319–335.

Stoehr, J., A. Benson, and N. Friel (2017). Noisy hamiltonian monte carlo for doubly-intractable distributions. *arXiv preprint arXiv:1706.10096*.

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of computational and graphical statistics 15*(2), 265–286.