Constraint Relaxation for Bayesian Modeling with Parameter Constraints

Leo Duan, Alexander L Young, 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. In this work, we first describe a general approach to utilize the large pool of unconstrained distributions in constrained space, then we propose to relax the parameter support into the neighborhood surrounding constrained space for convenient posterior estimation. The constraint relaxation can be done using data augmentation technique or with 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 sophisticated 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. Although MCMC is conceptually simple, except for a few limited cases, it is generally difficult to generate random variable strictly inside constrained space.

To overcome this difficulty, one common strategy is to reparameterize with un-/less constrained parameters at equal or less dimension. The new parameters form functions that can always satisfy the constraint. The transformation, if bijective, is known as 'coordinate system' in manifold embedding literature (Nash, 1954; Do Carmo, 2016). Examples include the polar coordinates for data on a hyper-sphere, or stick-breaking construction for Dirichlet distribution on probability simplex (Ishwaran and James, 2001). 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 tends to makes it more difficult to induce prior property on the original space. For example, uniformity of unconstrained parameter in a compact space may not be equivalent to uniformity on the constrained space via transformation. Diaconis et al. (2013) provide a useful tutorial and cautious guide on this subject.

Alternatively, it is typical to rely on customized solution 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, another drawback is that the tractable computation, especially posterior conjugacy, tends to break down under common modeling/data 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, which will create a low or zero acceptance probability. A recent 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. A third independent direction utilizes Hamiltonian Monte Carlo (HMC) that incorporates geometric structure with a Riemannian metric (Girolami and Calderhead, 2011), making proposals strictly inside the constrained space by solving a large linear system. Although simpler algorithms using geodesic flow were proposed for a few selected constrained space (Byrne and Girolami,

2013), compared to the first two strategies that operates in unconstrained space, strictly accommodating the constrained geometry tend to require more customization, such as computing the metric tensor for different manifolds.

The goal of this article is to dramatically expand the families of constrained priors one could use and develop simple computational strategy for more general constraints. We first introduce a general strategy to adapt common existing distributions into constrained space. To enable simple posterior computation, we *relax* the parameter into the neighborhood surrounding the constrained space. This approach enjoys the advantages of unconstrained space sampling while approximately takes into account of the geometry of the constrained space. This relaxation either produces an approximation for posterior under general constraints formed by equality and/or inequality, or an exact solution for several common constrained space such as simplex and Stiefel manifolds. Theoretic studies are conducted and comparison with existing approaches are shown in simulations and data applications.

2 Constrained Relaxation Methodology

Suppose θ is an $\mathcal{R}$-valued random variable with $\dim(\mathcal{R}) = r < \infty$ and that θ is subject to some constraints which restrict it to a subset $\mathcal{D} \subset \mathcal{R}$. In the Bayesian setting, of principle interest here, θ is a parameter where constraints arise from some given prior information. General approaches for sampling from the posterior in this case, are limited. Typically, it is necessary to find a reparameterization of θ or to build specialty sampling schemes. Thus, sampling θ from the constrained space, outside of some special cases is often difficult or computationally intractable.

One potential strategy to alleviate this issue is to construct an approximate distribution which places a high probability on $\mathcal{D}$ but has support in $\mathcal{R}$ by 'relaxing' the constraints. As a motivating example, consider the case where θ has density $\pi_{\mathcal{R}}(\theta)$ with support $\mathcal{R}$ and $\mathcal{D}$ is a measureable subset with positive measure. The posterior density of θ given data Y and $\theta \in \mathcal{D}$ is,

$$\pi(\theta|\theta \in \mathcal{D}, Y) \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{\mathcal{D}}(\theta)$$

with likelihood function $\mathcal{L}(\theta; Y)$, and data Y. As an approximation, suppose we used the density

$$\tilde{\pi}(\theta) \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp\left(-\frac{1}{\lambda} v_{\mathcal{D}}(\theta) \right) \tag{1}$$

where $v_{\mathcal{D}}(\theta) = \inf_{x \in \mathcal{D}} ||\theta - x||$ is a measure of the distance from θ to the constrained space for some metric $||\cdot||$.

Note that $\mathbb{1}_{\mathcal{D}}(\theta)$ is the pointwise limit of $\exp\big(-\nu_{\mathcal{D}}(\theta)/\lambda\big)$ (except perhaps on the boundary of $\mathcal{D}$) as $\lambda \to 0^+$. However, (1) has support $\mathcal{R}$ for all $\lambda > 0$, hence 'relaxing' the constraint. Since (1) is supported on

$\mathcal{R}$ it is more suitable for off-the-shelf MCMC sampling strategies. Ideally, one would hope that samples from (1) could be easily generated and that they would behave as if drawn from the fully conditioned distribution when λ is sufficiently small. We consider this approach when adapted to a number of settings, but generally we refer to it as constraint relaxation (**CORE**).

These observations motivate a number of questions about **CORE** which we investigate in the article. (i) For what types of distributions and constraints is CORE suitable? (ii) Is there a general approach for constructing the 'relaxed' constraint? (iii) How well do samples from the relaxed constraint represent those from the fully conditioned distribution? (iv) How does the approximation depend on the tuning parameter λ?

The answers to (ii) - (iv) depend largely upon (i). Therefore, beginning with (i), we assume θ is a continuous random variable (e.g. $\mathcal{R}$ is $\mathcal{R}$, $[0,\infty)^d$, $\mathbb{R}^{n\times k}$) and θ has an unconstrained distribution, $\pi_{\mathcal{R}}(\theta)$, which is absolutely continuous with respect to Lebesque measure on $\mathcal{R}$ hereby denoted as $\mu_{\mathcal{R}}$. We investigate two general types of constraints.

First, we consider the simpler case where $\mathcal{D}$ has positive measure, i.e. $\int_{\mathcal{D}} \pi(\theta)\mu_{\mathcal{R}}(d\theta) > 0$. Generally, inequality constraints (e.g. $a_i < \theta_i < b_i$, $a^T\theta < 0$, $||\theta||_2^2 < 1$) fall into this category. The analysis in this case will be more straightforward as we can follow traditional approaches to conditional probability. Here, the construction of the relaxed constraint will follow the form of Eq. (1).

Secondly, we consider the case where $\mathcal{D}$ is a measure zero subset of $\mathcal{R}$. In particular, we restrict ourselves to the setting where $\mathcal{D}$ can be represented implicitly as the solution set of a consistent system of equations $\{\nu_i(\theta) = 0\}_{i=1}^s$ so that $\mathcal{D} = \{\theta | v_j(\theta) = 0,\, j = 1,\ldots,s\}$ is a co-dimension s submanifold of $\mathcal{R}$. Additional requirement for the constraint functions are discussed in Section 3.2.

While these criteria may seem restrictive, we note that many common constraints (e.g. $||\theta||^2 = 1$, $\sum_i \theta_i = 1$, $\theta \in V_k(\mathbb{R}^n)$ where $V_k(\mathbb{R}^n)$ is the Stiefel manifold) fall into this category. In this case, the conditional distribution of θ given $\theta \in \mathcal{D}$, must be handled with care since $\int_D \pi_R(\theta) d\mu_{\mathcal{R}}(\theta) = 0$. However, the requirement that $\mathcal{D}$ has codimension s will serve two purposes. First, it will make the construction of conditional distributions on $\mathcal{D}$ using the tools of geometric measure theory more intuitive. Secondly, it will motivate a general strategy for choosing appropriate constraint equations and in constructing a relaxed density similar to (1).

Prior to discussing rigorous details on the use of **CORE**, we first review the principle details of the method in the context of a few examples. These examples were chosen to highlight the basic strategy of implementing **CORE** and to clarify some of the main ideas behind the supporting theory.

2.1 Positive measure constrained spaces

Consider the case where $\mathcal{D}$ has positive measure. The constrained posterior density, $\pi_{\mathcal{D}}$, for $\theta|\theta \in \mathcal{D}$ and data Y is

$$\pi_{\mathcal{D}}(\theta|Y) = \frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{\mathcal{D}}(\theta)}{\int_{\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)} \propto \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{\mathcal{D}}(\theta).$$

This constrained density is defined with respect to $\mu_{\mathcal{R}}$.

Suppose we approximate $\pi_{\mathcal{D}}$ with a relaxed density

$$\tilde{\pi}_\lambda(\theta) = \mathcal{L}(\theta;Y)\frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp\big(-\frac{v_d(\theta)}{\lambda}\big)}{\int_{\mathcal{R}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp\big(-\frac{v_d(\theta)}{\lambda}\big)d\mu_{\mathcal{R}}(\theta)} \propto \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp\big(-\frac{v_d(\theta)}{\lambda}\big)$$

which is also absolutely continuous with respect to $\mu_{\mathcal{R}}$. Here $\lambda > 0$ and $v_d(\theta)$ is a scalar-valued function which measures the distance from θ to the constrained space $\mathcal{D}$, i.e. $v_d(\theta) = 0\ \forall\theta \in \mathcal{D}$ and is positive otherwise. Formally, as $\lambda \to 0^+$, $\exp(-v_d(\theta)/\lambda) \to \mathbb{1}_{\mathcal{D}}(\theta)$ pointwise. If $\mathcal{D}$ is an open subset of $\mathcal{R}$, this limit may not hold on the boundary of $\mathcal{D}$, denoted $\partial\mathcal{D}$. However, in general $\mu_{\mathcal{R}}(\partial\mathcal{D}) = 0$ and we are working with densities. Thus, we can ignore this issue.

There are many possible choices for v_d which may be selected for different reasons. Perhaps the simplest choice is to take

$$v_d(z) = \inf_{x\in\mathcal{D}} ||z - x||_k \tag{2}$$

where $||\cdot||_k$ denotes the distance using the k-norm. Under this choice of v_d, the relaxation is isotrophic. More generally, one could use

$$v_d(z) = \inf_{x\in\mathcal{D}} \sqrt{(x-z)^T A(x-z)} \tag{3}$$

for some positive definite matrix A. In this case, the relaxation is anistrophic, and can be viewed as a form of directional relaxation. This choice of distance, v_d, allows for a more detailed specification of the rates at which individual components of θ relax to $\mathcal{D}$. In fact, if one seeks to constrain θ within a neighborhood containing $\mathcal{D}$, the matrix A can be chosen to preferentially relax the constraint and limit deviations from $\mathcal{D}$ in certain directions.

So long as $v_d(\theta)$ is zero for $\theta \in \mathcal{D}$ and positive for $\theta \in \mathcal{D}^c$, it follows that $\pi_{\mathcal{D}}$ is the pointwise limit of $\tilde{\pi}_\lambda$ for $\mu_{\mathcal{R}}$ a.e. θ in $\mathcal{R}$. It is natural to expect that estimates of $E[g(\theta)|\theta \in \mathcal{D}]$ can be obtained by using the relaxed posterior density, $\tilde{\pi}_\lambda$, rather than the fully constrained posterior density $\pi_{\mathcal{D}}$. For small λ, we anticipate

$$\int_{\mathcal{R}} g(\theta)\tilde{\pi}_\lambda(\theta)d\mu_{\mathcal{R}}(\theta) \approx \int_{\mathcal{D}} g(\theta)\pi_{\mathcal{D}}(\theta)d\mu_{\mathcal{R}}(\theta).$$

Rigorous details for the validity of this approximation, a suitable class of functions for which it applies, and rates of convergence in λ are contained in Section 3.1. For now, we discuss the implementation in the case of a bivariate normal distribution constrained to a triangular region in the $\theta_1 - \theta_2$ plane.

2.1.1 Example: Gaussian under Linear Inequality

Suppose that $\theta = (\theta_1, \theta_2)$ follows a bivariate Gaussian with mean $\mu \in \mathbb{R}^2$ and covariance matrix, $\Sigma = \sigma^2 I_2$, $\sigma > 0$, and that θ is constrained to the triangular region $\mathcal{D} = \{(\theta_1, \theta_2) | \, \theta_1 + \theta_2 \leq 1, \theta_1 \geq 0, \theta_2 \geq 0\}$. Let

$$v_d(\theta_1, \theta_2) = \max(\theta_1 + \theta_2 - 1, 0) + \max(-\theta_1, 0) + \max(-\theta_2, 0).$$

Then, $v_d(\theta_1, \theta_2) = 0 \; \forall \theta \in \mathcal{D}$. Otherwise, v_d is positive. The fully constrained density of θ given that $\theta \in \mathcal{D}$ is then,

$$\begin{aligned}
\pi_{\mathcal{D}}(\theta_1, \theta_2) = \pi(\theta_1, \theta_2 | \theta \in \mathcal{D}) &= \frac{\exp\left(- \frac{||\theta - \mu||^2}{2\sigma^2} \right) \mathbb{1}_{\mathcal{D}}(\theta)}{\int_0^1 \int_0^{1-\theta_1} \exp\left(- \frac{||\theta - \mu||^2}{2\sigma^2} \right) d\theta_2 d\theta_1} \\
&\propto \exp\left(- \frac{||\theta - \mu||^2}{2\sigma^2} \right) \mathbb{1}_{\{\theta_1 + \theta_2 \leq 1\} \cap \{\theta_1 \in [0, \infty)\} \cap \{\theta_2 \in [0, \infty)\}}.
\end{aligned}$$

Alternatively, the relaxed density, $\tilde{\pi}_\lambda(\theta)$, is

$$\begin{aligned}
\tilde{\pi}_\lambda(\theta) &= \frac{\exp\left(- \frac{||\theta - \mu||^2}{2\sigma^2} - \frac{1}{\lambda} v_d(\theta) \right)}{\int_{\mathbb{R}^2} \exp\left(- \frac{||\theta - \mu||^2}{2\sigma^2} - \frac{1}{\lambda} v_d(\theta) \right) d\theta_2 d\theta_1} \\
&\propto \exp\left(- \frac{||\theta - \mu||^2}{2\sigma^2} \right) \exp\left(- \frac{1}{\lambda} [\max(\theta_1 + \theta_2 - 1, 0) + \max(-\theta_1, 0) + \max(-\theta_2, 0)] \right).
\end{aligned} \tag{4}$$

Figure 1 depicts a few plots of the relaxed density as λ decreases. For $\lambda = 10^{-4}$ the constrained triangular region is clear. However, the relaxed density still places some small probability outside of the constrained region. This can be seen by observing the smooth decrease in the density near the boundary of the triangle.

In Section 5.1, we return to this example and use the relaxed density to sample.

2.2 Measure zero constraints

Alternatively, assume that $\mathcal{D}$ is a measure zero subset of $\mathcal{R}$, i.e. $\int_{\mathcal{D}} \mathcal{L}(\theta; Y) \pi_{\mathcal{R}} d\mu_{\mathcal{R}}(\theta) = 0$. In particular, we assume $\mathcal{D}$ is a submanifold of $\mathcal{R}$ such that $\dim(\mathcal{D}) = d =< r = \dim(\mathcal{R})$. The codimension of $\mathcal{D}$ is $s = r - d$. As such, the construction of the conditional distribution of $\theta | \theta \in \mathcal{D}$ must be handled carefully. One cannot simply re-normalize by a factor of $\left[\int_{\mathcal{D}} \mathcal{L}(\theta; Y) \pi_{\mathcal{R}} d\mu_{\mathcal{R}}(\theta) \right]^{-1}$. Instead, we propose to construct a *regular*

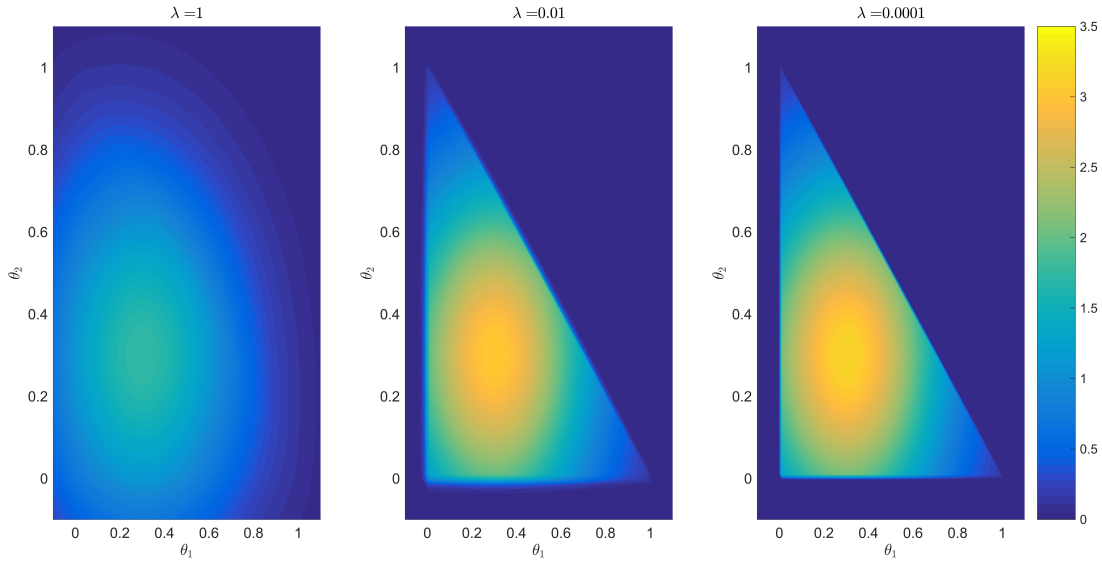


Figure 1: The relaxed distribution from Eq. (4) for decreasing is shown above when $\mu = [0.3, 0.3]^T$ and $\sigma^2 = 1/10$. For $\lambda = 1$, the distribution is similar to the unconstrained Gaussian. As λ decreases, the relaxation is reduced and the triangular constrained region becomes more apparent.

conditional probability (r.c.p.) which is consistent with unconstrained probability density $\pi_{\mathcal{R}}$. However, it is more natural to first discuss the construction of the relaxed density.

Similar to the approach of the previous section, we want a function $\nu_{\mathcal{D}}(\theta)$ which measures the distance from θ to $\mathcal{D}$. If $\theta \in \mathcal{D}$, then $\nu_{\mathcal{D}} = 0$. Otherwise, it is non-zero. For technical reasons, discussed in Section 3.2, $\nu_{\mathcal{D}}$ is a map from $\mathcal{R}$ to $\mathbb{R}^{r-d}$. Thus, $||\nu_{\mathcal{D}}(\theta)||_1 = 0$ implies that $\theta \in \mathcal{D}$ whereas, $||\nu_{\mathcal{D}}(\theta)||_1 > 0$ implies that $\theta \notin \mathcal{D}$. To construct the relaxed density we will multiply the unconstrained density by a factor of $\exp\left(-\frac{1}{\lambda}||\nu_{\mathcal{D}}(\theta)||\right)$. Therefore, the relaxed density, denoted $\tilde{\pi}_\lambda(\theta)$, is

$$\tilde{\pi}_\lambda(\theta) = \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)\exp\left(-\frac{1}{\lambda}||\nu_{\mathcal{D}}(\theta)||\right)}{\int_{\mathcal{R}}\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)\exp\left(-\frac{1}{\lambda}||\nu_{\mathcal{D}}(\theta)||\right)d\mu_{\mathcal{R}}(\theta)} \propto \mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)\exp\left(-\frac{1}{\lambda}||\nu_{\mathcal{D}}(\theta)||\right) \tag{5}$$

Much like the positive measure case, $\exp\left(-\frac{1}{\lambda}||\nu_{\mathcal{D}}(\theta)||\right)$ converges pointwise to $\mathbb{1}_{\mathcal{D}}(\theta)$. As $\lambda \to 0^+$, this multiplicative factor is concentrating the probability to a small layer around the constrained space.

A complete definition of the r.c.p. is given in the appendix. For now, the following intuition is sufficient. While $\mathcal{D}$ has zero volume (i.e. zero Lebesque measure), it will have a positive, possibly infinite d-dimensional 'surface area.' The r.c.p. allow us to construct a density with respect to this surface area which is consistent with the original unconstrained posterior. More rigorous details regarding the constrained space, the

construction of the regular conditional probability, and relevant convergence results are contained in Section 3.2. For the moment, we state the regular conditional probability

$$\pi_{\mathcal{D}}(\theta) = \pi(\theta|\theta \in \mathcal{D}) \propto \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu_{\mathcal{D}}(\theta))}\mathbb{1}_{\mathcal{D}}(\theta). \tag{6}$$

Here $J(\nu_{\mathcal{D}}(\theta)) = \sqrt{(D\nu_{\mathcal{D}})'(D\nu_{\mathcal{D}})}$ is the d-dimensional Jacobian of $\nu_{\mathcal{D}}$ which we assume is positive. The density is defined with respect to the normalized d-dimensional Hausdorff measure, $\bar{\mathcal{H}}^d$, which corresponds to d-dimensional surface area.

As a result, the expectation of $g(\theta)$ given $\theta \in \mathcal{D}$ is

$$E[g(\theta)|\theta \in \mathcal{D}] = \frac{\int_{\mathcal{D}} g(\theta)\frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu_{\mathcal{D}}(\theta))}d\bar{\mathcal{H}}^d(\theta)}{\int_{\mathcal{D}} \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu_{\mathcal{D}}(\theta))}d\bar{\mathcal{H}}^d(\theta)}.$$

Again, one can interpret the integration as a d-dimensional integral of surface area over $\mathcal{D}$.

Since the relaxed density places high probability on a shell surrounding $\mathcal{D}$, it is not unreasonable to expect that

$$E[g|\theta \in \mathcal{D}] = \int_{\mathcal{D}} g(\theta)\pi_{\mathcal{D}}d\bar{\mathcal{H}}^d(\theta) \approx \int_{\mathcal{R}} g(\theta)\tilde{\pi}_\lambda(\theta)d\mu_{\mathcal{R}}(\theta) = E_{\tilde{\Pi}}[g]$$

when λ is small. Like the positive measure case, we provide details for the validity of this approximation, a suitable class of functions for which it applies, and rates of convergence in λ..

2.2.1 Example: von Mises–Fisher on Unit Circle

Again, suppose that $\theta = (\theta_1, \theta_2)$ follows a bivariate Gaussian with mean $\mu \in \mathbb{R}^2$ and covariance matrix, $\Sigma = \sigma^2 I_2$, $\sigma > 0$ which is constrained to the unit circle $\mathcal{D} = \{(\theta_1, \theta_2)|\, \theta_1^2 + \theta_2^2 = 1\}$. Since the unit circle is one-dimensional and $\theta = (\theta_1, \theta_2)$ is two-dimensional, the constraint function must be (2-1)=1-dimensional. Let

$$v_d(\theta_1, \theta_2) = \theta_1^2 + \theta_2^2 - 1.$$

Then, $v_d(\theta_1, \theta_2) = 0\ \forall \theta \in \mathcal{D}$. Otherwise, v_d is non-zero. Furthermore, $J(\nu_{\mathcal{D}}(\theta)) = 2||\theta_||_2 = 2$ for $\theta \in \mathcal{D}$. The fully constrained density of θ given that $\theta \in \mathcal{D}$ is then,

$$\pi_{\mathcal{D}}(\theta_1, \theta_2) = \pi(\theta_1, \theta_2|\theta \in \mathcal{D}) = \frac{\exp\left(-\frac{||\theta-\mu||^2}{2\sigma^2}\right)\mathbb{1}_{\mathcal{D}}(\theta)}{\int_{\mathcal{D}} \exp\left(-\frac{||\theta-\mu||^2}{2\sigma^2}\right)d\bar{\mathcal{H}}^1(\theta)}$$

$$\propto \exp\left(-\frac{||\theta-\mu||^2}{2\sigma^2}\right)\mathbb{1}_{\theta_2^1+\theta_2^2=1}.$$

This density should be interpreted with respect to the normalized Hausdorff-1 measure on the unit circle which coincides with arclength in this case.

It is more natural to parameterize the circle $(\theta_1, \theta_2) = (\cos\phi, \sin\phi)$ for $\phi \in [0, 2\pi)$. The density of angle ϕ is

$$\pi_\phi(\phi) = \pi_{\mathcal{D}}(\cos\phi, \sin\phi) \propto \exp\left(\frac{r}{\sigma^2}\cos(\phi - \varphi)\right).$$

where $\mu = r(\cos\varphi, \sin\varphi)$. One should observe that this is the von-Mises–Fisher distribution on the unit circle with location φ and concentration r/σ^2.

Alternatively, the relaxed density, $\tilde{\pi}_\lambda(\theta)$, is

$$\begin{aligned}\tilde{\pi}_\lambda(\theta) &= \frac{\exp\left(-\frac{||\theta-\mu||^2}{2\sigma^2} - \frac{1}{\lambda}|v_d(\theta)|\right)}{\int_{\mathbb{R}^2}\exp\left(-\frac{||\theta-\mu||^2}{2\sigma^2} - \frac{1}{\lambda}|v_d(\theta)|\right)d\theta_2 d\theta_1} \\ &\propto \exp\left(-\frac{||\theta-\mu||^2}{2\sigma^2}\right)\exp\left(-\frac{1}{\lambda}|\theta_1^1 + \theta_2^2 - 1|\right).\end{aligned} \tag{7}$$

Figure 2 depicts a few plots of the relaxed density as λ decreases. For $\lambda = 10^{-2}$ the constraint along the circle is clear. While the relaxed density still places some small probability outside of the constrained region, the rightmost plot looks similar to the von-Mises distribution on the circle plotted in two dimensions.

An analogous example in $\mathbb{R}^3$ is considered in Section 5, in which we use **CORE** to generate samples from a von-Mises–Fisher distribution on S^2. Furthermore, we show that **CORE** can be used to generalize the von-Mises–Fisher distribution by constraining a multivariate t-distribution to the sphere.

3 Analytical Results

3.1 Positive measure constrained spaces

We return now to the case where $\mathcal{D}$ has positive measure with respect to the unconstrained posterior density so

$$\int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) > 0.$$

Recall that in this case, the fully constrained posterior distribution is

$$\pi_{\mathcal{D}}(\theta|Y) = \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{\mathcal{D}}(\theta)}{\int_{\mathcal{D}}\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)} \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{\mathcal{D}}(\theta).$$

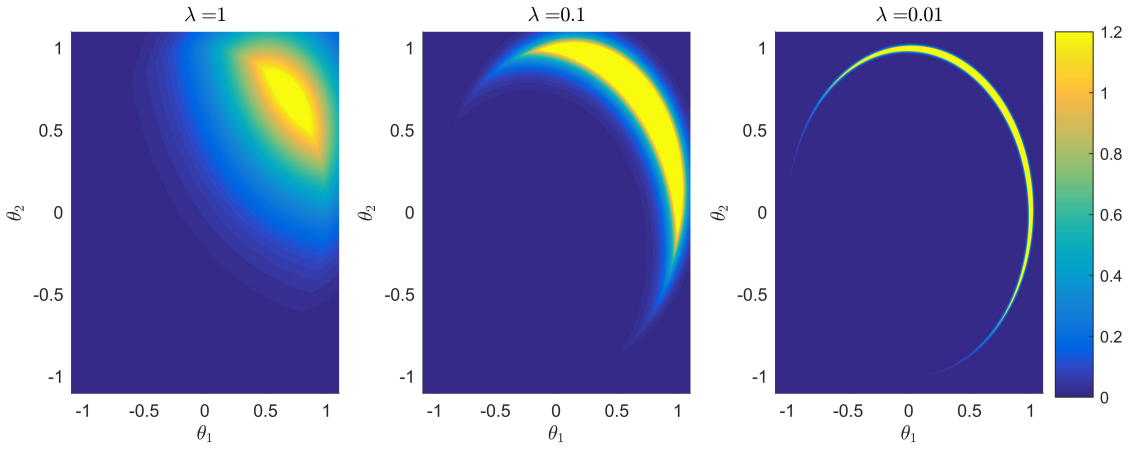


Figure 2: The relaxed distribution from Eq. (7) for decreasing λ is shown above when $\mu = (1/\sqrt{2}, 1/\sqrt{2})$ and $\sigma^2 = 1/25$. This corresponds to a von-Mises distribution with location $\pi/4$ and concentration $1/25$. As λ decreases, the relaxation is reduced and the circular constraint becomes clear.

Whereas, the relaxed posterior density is

$$\tilde{\pi}_\lambda(\theta) = \mathcal{L}(\theta; Y) \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp\big(-\frac{v_d(\theta)}{\lambda}\big)}{\int_{\mathcal{R}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp\big(-\frac{v_d(\theta)}{\lambda}\big) d\mu_{\mathcal{R}}(\theta)} \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp\big(-\frac{v_d(\theta)}{\lambda}\big)$$

for some distance function $v_d(\theta)$ which is zero for $\theta \in \mathcal{D}$ and positive for $\theta \in (\mathcal{D}^c)^o$.

Both of these densities are absolutely continuous with respect to Lebesque measure on $\mathcal{R}$. Thus, the expectation of g given $\theta \in \mathcal{D}$, is

$$E[g(\theta)|\theta \in \mathcal{D}] = \int_{\mathcal{D}} g(\theta)\pi_{\mathcal{D}}(\theta) d\mu_{\mathcal{R}} = \frac{\int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) d\mu_{\mathcal{R}}(\theta)}{\int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) d\mu_{\mathcal{R}}(\theta)}. \tag{8}$$

Similarly, the expected value of g with respect to the relaxed density, denote as $E_{\tilde{\pi}_\lambda}[g]$, is then,

$$E_{\tilde{\pi}_\lambda}[g(\theta)] = \int_{\mathcal{R}} g(\theta)\tilde{\pi}_\lambda(\theta) d\mu_{\mathcal{R}} = \frac{\int_{\mathcal{R}} g(\theta)\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-v_d(\theta)/\lambda) d\mu_{\mathcal{R}}(\theta)}{\int_{\mathcal{R}} \mathcal{L}(\theta; Y) \exp(-v_d(\theta)/\lambda)\pi_{\mathcal{R}}(\theta) d\mu_{\mathcal{R}}(\theta)}. \tag{9}$$

We can now consider the behavior of $E_{\tilde{\pi}_\lambda}[g]$ as $\lambda \to 0^+$.

Lemma 1. *Suppose $g \in \mathbb{L}^1(\mathcal{R}, \mathcal{L}(\theta; Y)\pi_{\mathcal{R}} d\mu_{\mathcal{R}})$. Then,*

$$\Big| E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)] \Big| \leq \frac{\int_{\mathcal{R}\setminus\mathcal{D}} (C_{\mathcal{R}} E|g(\theta)| + |g(\theta)|)\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-v_d(\theta)/\lambda) d\mu_{\mathcal{R}}(\theta)}{\big[\int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) d\mu_{\mathcal{R}}(\theta)\big]^2}$$

where $E|g(\theta)| \propto \int_{\mathcal{R}} |g(\theta)|\mathcal{L}(\theta; Y)\pi_{\mathcal{R}} d\mu_{\mathcal{R}}(\theta)$ is the expected value of $|g(\theta)|$ with respect to the unconstrained posterior density and $C_{\mathcal{R}} = \int_{\mathcal{R}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) d\mu_{\mathcal{R}}(\theta)$ is the normalizing constant of this unconstrained posterior density. Furthermore, if $v_d(\theta)$ is zero for all $\theta \in \mathcal{D}$ and positive for $\theta \in (\mathcal{R} \setminus \mathcal{D})^o$, it follows from the

dominated convergence theorem that

$$\left| E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)] \right| \to 0 \text{ as } \lambda \to 0^+.$$

Thus, one can obtain sufficiently accurate estimates of $E[g|\theta \in \mathcal{D}]$ by sampling from $\tilde{\pi}_\lambda$ when λ is sufficiently small. From a practical standpoint, it is desirable to understand the rate at which $E_{\tilde{\pi}_\lambda}[g(\theta)]$ converges to $E[g(\theta) \in \mathcal{D}]$. This question is addressed in the following theorem.

Theorem 1. *Suppose $g \in \mathbb{L}^2(\mathcal{R}, \mathcal{L}(\theta;Y)\pi_{\mathcal{R}} d\mu_{\mathcal{R}})$, $v_d(\theta)$ has the form of Eq. (2) with $k = 2$, $\mathcal{D}$ has a piecewise smooth boundary, and that $\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)$ is continuous on a open neighborhood containing $\mathcal{D}$. Then for $0 < \lambda \ll 1$,*

$$\left| E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)] \right| = O(\sqrt{\lambda}).$$

This theorem follows by applying the Cauchy-Schwartz inequality to the term in the numerator of the bound given in Lemma 1. One can attain a bound depending on the surface area of $\mathcal{D}$ when it is bounded. More generally, if $\mathcal{D}$ is a unbounded subset of $\mathcal{R}$, one can make use of the co-area formula Federer (2014); Diaconis et al. (2013) which is stated in the next section. The proofs of Lemma 1 and Theorem 1 are contained in Appendix A.

These results have some important implications both analytically and numerically. First, in addition to point estimates, $E[\theta|\theta \in \mathcal{D}]$, it is possible to approximate probabilities $P(\theta \in \mathcal{F}|\theta \in \mathcal{D})$ and higher moments, e.g. $E[\Pi_j \theta_j^{k_j}|\theta \in \mathcal{D}]$, so long as these moments exist for the unconstrained posterior density.

Secondly, these bounds demonstrate that the error in using the relaxed density to approximate $E[g(\theta)|\theta \in \mathcal{D}]$ is proportional to $\sqrt{\lambda}[\int_{\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)]^{-2}$ although this rate may not be optimal. In practice, λ may need to be very small, particularly in the case where $0 < P(\theta \in \mathcal{D}) \ll 1$. Of course, specific details of the scaling of $\left| E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)] \right|$ will depend upon the choice of $\mathcal{D}$ and $v_d(\theta)$. In practice (see Sections 5), we have observed much better convergence rates.

In general, one avenue for mitigating numerical difficulties which may arise when $\lambda \ll 1$ is needed is to use Eq. 3 to relax the density in directions where accuracy is less important. Fortunately, both Lemma 1 and Theorem 1 will still hold when A is positive definite since this corresponds to a linear re-scaling of the parameter space.

3.2 Measure zero constrained spaces

Prior to investigating the application of **CORE** in the measure zero case, we review a few important concepts of geometric measure theory which are used throughout this section. First, recall the definition of Hausdorff

measure.

Definition - Hausdorff Measure. *Let $A \subset \mathcal{R}^r$. Fix $s \leq r$. Then*

$$\mathcal{H}^s(A) = \lim_{\delta \to 0} \inf \left\{ \sum [diam(S_i)]^s : A \subseteq \cup S_i, diam(S_i) \leq \delta, diam(S_i) = \sup_{x,y \in S} \|x - y\| \right\}$$

.

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 $s = r$, Lebesgue and normalized Hausdorff measures coincide $\mu_{\mathbb{R}^m}(A) = \bar{\mathcal{H}}^s(A)$ (Evans and Gariepy, 2015). Additionally, for a subset $\mathcal{D}$, there exists a unique, critical value d such that

$$\bar{\mathcal{H}}^s(\mathcal{D}) = \begin{cases} 0, & s > d \\ \infty, & s < d. \end{cases}$$

The critical value, d, is referred to as the Hausdorff dimension of $\mathcal{D}$. We note that, when $\mathcal{D}$ is a compact, d-dimensional submanifold of $\mathbb{R}^m$, it will have Hausdorff dimension d and $\bar{\mathcal{H}}^d(\mathcal{D})$ is the d-dimensional surface area of A.

We now state the co-area formula which is used to define a regular conditional probability on the measure zero constrained space $\mathcal{D}$ and is pivotal in all of the proofs of the theorems.

Theorem 2. *Co-area formula (Diaconis et al., 2013; Federer, 2014) Suppose $\nu : \mathbb{R}^r \to \mathbb{R}^s$ with $s < r$ is Lipschitz and that $g \in \mathbb{L}^1(\mathbb{R}^r, \mu_{\mathbb{R}^r})$. Assume $J(\nu(\theta)) > 0$, then*

$$\int_{\mathbb{R}^r} g(\theta) J\nu(\theta) d\mu_{\mathbb{R}^r}(\theta) = \int_{\mathbb{R}^s} \left(\int_{\nu^{-1}(y)} g(\theta) d\bar{\mathcal{H}}^{r-s}(\theta) \right) d\mu_{\mathbb{R}^s}(y), \tag{10}$$

The behavior of the pre-images $\nu^{-1}(y)$ in the co-area formula are important for the convergence results presented later in this section. As such, we require the following properties of $\mathcal{D}$.

Hereafter, we assume that $\mathcal{D}$ can be defined implicitly as the solution set to a system of s equations, $\{\nu_j(\theta) = 0\}_{j=1}^s$, where

(a) $\nu_j : \mathcal{R} \to \mathbb{R}$ is Lipschitz continuous,

(b) $v_j(\theta) = 0$ only for $\theta \in \mathcal{D}$,

(c) for $k = 1, \ldots, s$, the preimage $v_k^{(-1)}(x)$ is a co-dimension 1 sub-manifold of $\mathcal{R}$ for $\mu_{\mathbb{R}}$-a.e. x in the range of ν_k,

(d) $\nu_j^{(-1)}(0)$ and $\nu_k^{(-1)}(0)$ intersect transversally for $1 \leq j < k \leq s$.

We refer to the functions $\nu_1, \ldots, \nu_s$ as constraint functions. In this case, if we let $\nu : \mathcal{R} \to \mathbb{R}^s$ be the vector-valued function $\nu(\theta) = [\nu_1(\theta), \ldots, \nu_s(\theta)]^T$, then $\mathcal{D} = \ker(v)$ is a co-dimension s submanifold of $\mathcal{R}$ for $\mu_{\mathbb{R}^s}$-a.e. x the range of v. Recall, the ambient space, $\mathcal{R}$, is $r-$dimensional. Therefore, it follows that $\mathcal{D}$ is a $(r-s)-$dimensional submanifold of $\mathcal{R}$, and it is natural to discuss the $(r-s)$-dimensional surface area of $\mathcal{D}$.

Property (a), guarantees that ν is itself Lipschitz. The remaining properties (b)-(d) are constructed so that $\nu^{(-1)}(x)$ for $x \in \mathbb{R}^s$ is also a submanifold which is close to $\mathcal{D} = \nu^{(-1)}(0)$ when x is near zero. In particular, the assumption of transversality ensures that $\nu^{(-1)}(x)$ with also be $r-s$ dimensional for x sufficiently close to 0.

The existence and uniqueness of the constraints must be addressed. In the case where $\mathcal{D}$ is specified by a collection of equality constraints – such as the probability simplex or the Stiefel manifold for example– it is not difficult to find a suitable set of constraint functions. Table 1 contains a number of examples of common constrained spaces and appropriate choices of constraint functions.

$\mathcal{R}$	$\mathcal{D}$	$\dim(R)$	$\dim(D)$	Constraint functions
$[0,1]^r$	Probability simplex, Λ	r	$r-1$	$\nu(\theta) = \sum(\theta) - 1$
$\mathbb{R}^r$	Line, span$\{\vec{u}\}$ $\vec{u} \neq \vec{0}$	r	1	$\nu_j(\vec{\theta}) = \vec{\theta}^T \vec{b}_j$ $\{\vec{b}_1, \ldots, \vec{b}_{r-1}\}$ a basis for span$\{\vec{u}\}^\perp$
$\mathbb{R}^r$	Unit sphere, $\mathbb{S}^{r-1}$	r	$r-1$	$\nu(\theta) = \arctan(\|\|\theta\|\|^2 - 1)$
$\mathbb{R}^{n\times k}$	Stiefel manifold, $V_k(\mathbb{R}^n)$ $\theta = [\vec{\theta}_1 \| \ldots \| \vec{\theta}_k]$, $\vec{\theta}_j \in \mathbb{R}^n$	nk	$nk - \frac{1}{2}k(k+1)$	$\nu_{i,j}(\theta) = \arctan(\vec{\theta}_i' \vec{\theta}_j - \delta_{i,j})$ $1 \leq i \leq j \leq k$ and $\delta_{i,j} = \mathbb{1}_{i=j}$

Table 1: Table of constraints for some commonly used constrained spaces.

In the more difficult situation where equality constraints are not given, finding $\{\nu_j\}_{j=1}^s$ may be very challenging. For the moment, we will assume that one can construct sufficiently accurate numerical approximations perhaps through cubic-splines or Fourier series, so that we may ignore the issue of existence.

With regards to uniqueness, we note that the constraints cannot be unique in any case. For example, given any constraints $\{\nu_j\}_{j=1}^s$ which satisfy (a)-(d) and non-zero constants $\{\lambda_j\}_{j=1}^s$, the constraints $\{\lambda_j \nu_j\}_{j=1}^s$ will also satisfy (a)-(d). It is then natural to wonder if one can find an optimal choice of the constraints. An optimal choice will depend largely on the properties of the constrained distribution that one wishes to estimate making the choice of $\{\nu_j\}_{j=1}^s$ context dependent. As such, we will address this issue in the examples contained in later sections. For now, let us proceed assuming that a suitable choice of $\{\nu_j(\theta)\}_{j=1}^s$ has been made.

Under the given construction of the constrained space, we can now specify the regular conditional probability of θ, given $\theta \in \mathcal{D}$.

Theorem 3. ? *Assume that $J(v(\theta)) > 0$ and that for each $z \in \mathbb{R}^s$ there is a finite non-negative integer p_z such that,*

$$m^{p_z}(z) = \int_{\mathbb{R}^s} \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=z}}{J(v(\theta))} d\bar{\mathcal{H}}^p(z) \in (0, \infty).$$

Then, for any Borel subset, E, of $\mathcal{R}$, it follows that

$$P(E|v(\theta) = z) = \begin{cases} \frac{1}{m^{p_z}(z)} \int_E \frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=z}}{J(v(\theta))} d\bar{\mathcal{H}}^p(z) & m^s(z) \in (0, \infty) \\ \delta(E) & m^p(z) \in \{0, \infty\} \end{cases}$$

is a valid regular conditional probability for $\theta \in \mathcal{D}$. Here, $\delta(E) = 1$ if $0 \in E$ and 0 otherwise.

By construction, $\{\theta : v(\theta) = z\}$ is a $(r - s)$ dimensional submanifold of $\mathcal{R}$ for $\mu_{\mathbb{R}^s}$-a.e. z in the range of ν. As such, it follows that one should take $p_z = r - s$. To reiterate, when $\mathcal{D}$ is contained in a compact subset of $\mathcal{R}$, which is the case for the unit circle and the Stiefel manifold, it follows that $m^{r-s}(z) \in (0, \infty)$ for $\mu_{\mathbb{R}^s}$-a.e. z in the range of ν. It is possible that $m^p(z) \in \{0, \infty\}$ for $p = 1, \ldots, r - 1$. For example, if $\mathcal{D}$ is an unbounded subset of $\mathcal{R}$ and $\frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=z}}{J(v(\theta))}$ decays sufficiently slowly, then $m^{r-s}(z) = \infty$. See Diaconis et al. (2013) for additional discussion of this issue. However, in most practical applications Theorem 3 allows us to define

$$\pi_{\mathcal{D}}(\theta|\theta \in \mathcal{D}, Y) = \frac{1}{m^{r-s}(\mathbf{0})} \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=\mathbf{0}}}{J(v(\theta))} \tag{11}$$

as the fully constrained posterior density as long as $\pi_{\mathcal{R}}$ decays sufficiently fast when $\mathcal{D}$ is an unbounded subset of $\mathcal{R}$.

Note that π_D is absolutely continuous with respect to the $(r - s)$-dimensional Hausdorff measure on $\mathcal{D}$ in the sense that

$$P(\theta \in F|\theta \in \mathcal{D}, Y) = \int_F \frac{1}{m^{r-s}(\mathbf{0})} \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=\mathbf{0}}}{J(v(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta)$$

for all measurable sets $\mathcal{F} \subset \mathcal{R}$. As a result, we can define the conditional expectation of $g(\theta)$ given $\theta \in \mathcal{D}$ as

$$E[g(\theta)|\theta \in \mathcal{D}] = E[g(\theta)|\nu(\theta) = 0\,] = \int_{\mathcal{R}} g(\theta)\pi_{\mathcal{D}}(\theta) d\bar{\mathcal{H}}^{r-s}(\theta).$$

Alternatively, the expected value of $g(\theta)$ with respect to the relaxed density, denote $E_{\tilde{\Pi}}[g(\theta)]$, is

$$\begin{aligned} E_{\tilde{\Pi}}[g(\theta)] &= \frac{\int_{\mathcal{R}} g(\theta)\pi_{\mathcal{R}}(\theta)\mathcal{L}(Y; \theta) \exp\left(-\frac{1}{\lambda}||\nu(\theta)||_1 \right) d\mu_{\mathcal{R}}(\theta)}{\int_{\mathcal{R}} \pi_{\mathcal{R}}(\theta)\mathcal{L}(Y; \theta) d\mu_{\mathcal{R}}(\theta)} \\ &= \frac{1}{m_\lambda} \int_{\mathcal{R}} g(\theta)\pi_{\mathcal{R}}(\theta)\mathcal{L}(Y; \theta) \exp\left(-\frac{1}{\lambda}||\nu(\theta)||_1 \right) d\mu_{\mathcal{R}}(\theta) \end{aligned} \tag{12}$$

where

$$m_\lambda = \int_{\mathcal{R}} g(\theta)\pi_{\mathcal{R}}(\theta)\mathcal{L}(Y;\theta)\exp\left(-\frac{1}{\lambda}||\nu(\theta)||_1\right)d\mu_{\mathcal{R}}(\theta).$$

The primary results of the section are the following statements regarding the use of $E_{\tilde{\Pi}}[g]$ to estimate $E[g|\theta \in \mathcal{D}]$.

Theorem 4. *Let $m : \mathbb{R}^s \to \mathbb{R}$ and $G\mathbb{R}^s \to \mathbb{R}$ be defined as follows*

$$m(x) = \int_{\nu^{-1}(x)} \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta)}{J(\nu(\theta))} d\mathcal{R}^{r-s}(\theta)$$

$$G(x) = \int_{\nu^{-1}(x)} g(\theta)\frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta)}{J(\nu(\theta))} d\mathcal{R}^{r-s}(\theta).$$

Suppose that both m and G are continuous on an open interval containing the origin and that $g \in \mathbb{L}^1(\mathcal{R}, \pi_R\mathcal{L}(y;\theta)d\mu_{\mathcal{R}})$. Then,

$$\left|E_{\tilde{\Pi}}[g] - E[g|\theta \in \mathcal{D}]\right| \to 0 \text{ as } \lambda \to 0^+.$$

Corollary 1. *In addition to the assumptions of Theorem 4, suppose that both m and G are differentiable at 0. Then*

$$\left|E_{\tilde{\Pi}}[g] - E[g|\theta \in \mathcal{D}]\right| = O\left(\frac{\lambda}{|\log\lambda|^s}\right)$$

as $\lambda \to 0^+$.

Like the results from Section 3.1, the convergence rates are sub-linear. However, unlike the positive measure case, the convergence rates are dimension dependent. Again, one can alter the individual constraints ν_j to alleviate computational demands if small values of λ are needed to attained desired accuracy when estimating $E[g|\theta \in \mathcal{D}]$. These issues are discussed further in the Examples contained in Section 5. For now, we review details regarding sampling algorithms which can be used in conjunction with **CORE**.

4 Posterior Computation

Adapting unconstrained density into space $\mathcal{D}$ often disrupts its posterior conjugacy. Since one can now sample the posterior in $\mathcal{R}$ using CORE, 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 updating of parameters.

4.1 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-CORE (**??**), and $\{\theta^*, w\}$ under DA-CORE (**??**).

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,

$$\begin{aligned} &H(q,p) = U(q) + K(p), \\ &\text{where } U(q) = -\log \pi(q), \\ &\qquad K(p) = \frac{p' M^{-1} p}{2}. \end{aligned} \tag{13}$$

with $\pi(q)$ is the unnormalized density in (**??**) or (**??**).

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:

$$\begin{aligned} \frac{\partial q^{(t)}}{\partial t} &= \frac{\partial H(q,p)}{\partial p} = M^{-1} p, \\ \frac{\partial p^{(t)}}{\partial t} &= -\frac{\partial H(q,p)}{\partial q} = -\frac{\partial U(q)}{\partial q}. \end{aligned} \tag{14}$$

The exact solution to (14) is typically intractable but a valid Metropolis proposal can be generated by numerically approximating (14) 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:

$$p \leftarrow p - \frac{\epsilon}{2}\frac{\partial U}{\partial q}, \quad q \leftarrow q + \epsilon M^{-1} p, \quad p \leftarrow p - \frac{\epsilon}{2}\frac{\partial U}{\partial q} \tag{15}$$

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

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

4.2 Computing Efficiency and Support Expansion

Since CORE 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 quantification of the effects and provide a practical guidance on choosing $\pi(w)$ or λ in the two 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 via support expansion.

For approximation-CORE (**??**), to control approximation error, one can choose to relax a subset of constringent constraints. Observing $\mathcal{D} = \cap_k D_k$, each approximation $\exp(-\frac{|v_k(\theta^*)|^\alpha}{\lambda_k})$ corresponds to a constrained space $\mathcal{D}_k$. One practical strategy is that, for $\mathcal{D}_k$'s with $\eta(\theta; \mathcal{D}_k) \approx 0$, one uses moderate λ_k to induce some support expansion (denoted by $\lambda_k \geq \zeta$ with ζ moderately small but not too 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'. Noting this could create inaccuracy of HMC near the boundary with $\lambda_k \approx 0$ under fixed step size, we use random step size ϵ at each iteration to reduce the error.

The Hessian $\mathbf{H}_U(q)$ under approximation-CORE is given by

$$\mathbf{H}_U(q) = -\mathbf{H}_{\log L(y;\theta^*)\pi_{\mathcal{R}}(\theta^*)/J(v(\theta^*))}(\theta^*) + \sum_k \lambda_k^{-1}\mathbf{H}_{|v_k|^\alpha}(\theta^*)\mathbb{1}_{\theta\notin\mathcal{D}_k}, \tag{16}$$

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 the 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, as quantified above, the approximation error is often $\mathcal{O}(\max_{\lambda_k\geq\zeta}\lambda_k)$ and decreases faster than the efficiceny cap $\mathcal{O}(\min_{\lambda_k\geq\zeta}\lambda_k^{1/2})$, as λ_k decreases. In our experiments, 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 DA-CORE (**??**) does not involve such error trade-off, it is preferred when it is applicable. Letting $\mathcal{Q}_\theta \subset \mathcal{D}$ denote the support for the constrained $\theta \in \mathcal{D}$, the reparameterization changes the support to $Q_{\theta^*} = \{g(\theta; w) : \theta \in \mathcal{Q}\}$. Therefore, one could choose $\pi(w)$ to substantially increase $\eta(\theta^*; Q_{\theta^*})$. Since the augmented variable w is redundant, DA-CORE can be considered as one type of parameter expansion discovered by Liu and Wu (1999), who originally focused on accelerating Gibbs sampling of probit regression. Although for greater space expansion, it is possible to use diffuse or even improper prior for $\pi(w)$ on Q_{θ^*}, we recommend assigning $\pi(w)$ loosely centered at $w_0 : g(\theta; w_0) = \theta$ (corresponding to when $\theta^* = \theta$), which makes θ^* a mild relaxation of θ. This ensures no substantial change in $\xi_1^{-1/2}(\theta^*)$ in HMC.

5 Simulations

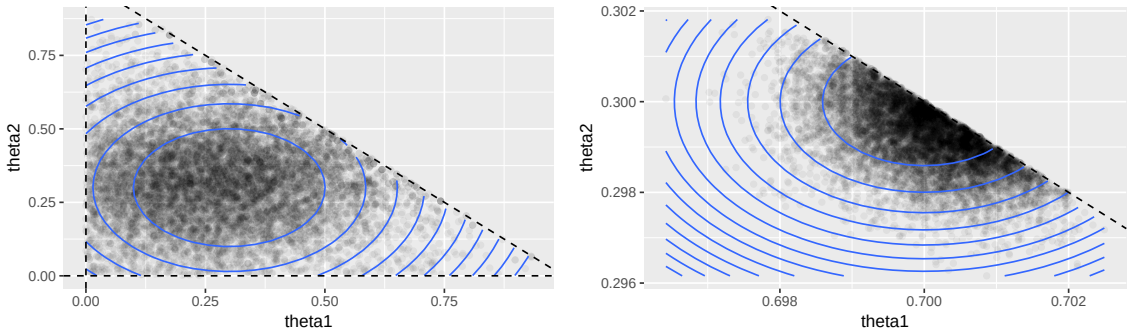
In order to compare against existing approaches on computing efficiency and provide empirical evidence supporting our previous result, we run simulations on several toy examples in this section.

5.1 Example: Gaussian under Linear Inequality

In this example, we consider a linear model under a system of linear inequality constraints. Although a recent work proposed a new customized prior with exhibits posterior conjugacy (Danaher et al., 2012), via the **CORE** framework, one can make use of a general Gaussian prior.

Suppose that we wish to sample from a bivariate Gaussian $\theta = (\theta_1, \theta_2) \sim \text{No}\left(\mu, I\sigma^2\right)$ where $\mathcal{D} = \{(\theta_1, \theta_2)|\theta \in (0,1)^2,\ \theta_1 + \theta_2 < 1\}$, which forms a triangle in the $\theta_1 - \theta_2$ plane. We investigate this example for two different choices of (μ, σ^2). First, let $(\mu, \sigma^2) = ([0.3, 0.3], 1/10)$ which centers the distribution of θ near the center of $\mathcal{D}$. Secondly, $(\mu, \sigma^2) = ([0.7, 0.3]', 1/10^4)$ which induces a distribution which is concentrated on the boundary of $\mathcal{D}$.

Let $v_d(\theta) = |1 - \theta_1 + \theta_2|_+ + |-\theta_1|_+ + |-\theta_2|_+ + |1-\theta_1|_+ + |1-\theta_2|$ where $|x|_+ = \max(x, 0)$. Since the triangle has wide support with $\eta(\theta; \mathcal{D})$ away from 0, small $\lambda = 10^{-8}$ guarantees almost no approximation error. Figure 3 contains plots which show posterior samples and contours of the constrained posterior density. Clearly, the posterior samples are all inside $\mathcal{D}$. To compare, we ran simple rejection sampling with untruncated normal proposal $\text{No}(\mu, I\sigma^2)$. As expected, 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 3: 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.

5.2 Example: von Mises–Fisher on Unit Circle

To illustrate equality constraint relaxation, 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-CORE using $\exp(-\frac{|\theta'\theta - 1|}{\lambda})$ for approximating the indicator, DA-CORE using $\theta_1 = \frac{\theta_1^*}{w}, w = \sqrt{(\theta_1^*)^2 + (\theta_2^*)^2}$ and $\pi(w) \sim \text{No}(1, 1)\mathbb{1}_{w>0}$ and exact von Mises–Fisher obtained using 'movMF' package.

Unlike the previous linear inequality constraint, the unit circle has narrow $\eta(\theta; \mathcal{D}) = 0$ 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-CORE. 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 2 shows the effective sample size per 1000 iterations, the effective 'violation' $|v(\theta)| = |\theta_1^2 + \theta_2^2 - 1|$ 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. As expected, the DA-CORE is exact

and has high effective sample size.

	HMC based on CORE				Exact
	Approximation			DA-CORE	
	$\lambda = 10^{-3}$	$\lambda = 10^{-4}$	$\lambda = 10^{-5}$		
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 2: Benchmark of constraint relaxation methods on sampling von–Mises Fisher distribution on a unit circle. For each approximation CORE, 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 DA-CORE and approximation-CORE with relatively large λ have high computing efficiency.

5.3 Dirichlet on a Simplex

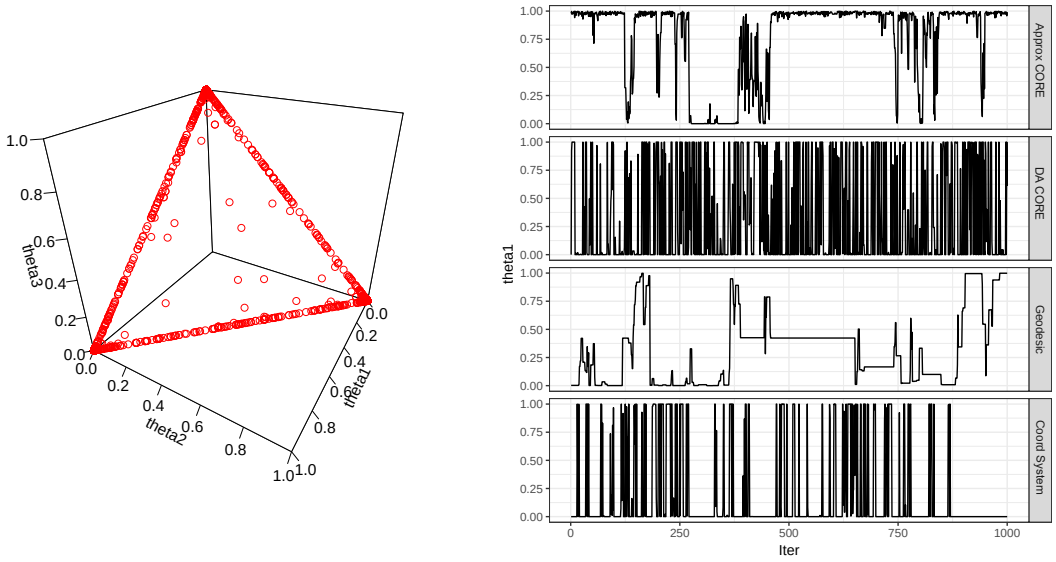
Lastly, we experiment with a particularly challenging distribution on a $(p-1)$-simplex, defined by $\{\theta : \theta \in (0,\infty)^p, \sum_{i=1}^p \theta_i = 1\}$. We consider Dirichlet distribution $\mathrm{Dir}(\alpha)$, with $\pi_{\mathcal{D}}(\theta) \propto \prod_{i=1}^p \theta_i^{\alpha-1}$. When the concentration parameter $\alpha < 1$, $\mathrm{Dir}(\alpha)$ exhibits sparse property that some θ_i's become very close to 0, which is exploited in topic modeling (Wang and Blei, 2009) and shrinkage (Bhattacharya et al., 2015) literature. 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 4(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 approximation-CORE and DA-CORE. 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 (Ishwaran and James, 2001); 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 to have effective sample size as large as possible.

Table 3 lists the effective sample sizes under different α's. As α becomes smaller than 1, approximation-CORE and geodesic HMC become worse in performance, while DA-CORE and coordinate system are much less impacted. Figure 4(b) shows at $\alpha = 0.01$, the approximation-CORE and geodesic HMC are stuck for a long time, while DA-CORE works substantially better. As a well-tested reparameterization, HMC based on coordinate system still works acceptably well in this case.

The difficulty that approximation-CORE encountered was anticipated. Byrne and Girolami (2013) have previously reported similar slow-down of geodesic HMC computing on hyper-Dirichlet distribution (Hankin et al., 2010) with $\alpha < 1$. Comparing these two approaches, geodesic HMC relies on restricting the kinetic

flow on $\mathcal{D}$ via its product with the metric tensor, and approximation-CORE relies on creating high energy wall in the potential energy. The latter can be viewed as an approximation to the former, which explains the similarity in performance.



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

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

Figure 4: 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-CORE with $\lambda = 10^{-3}$, DA-CORE, geodesic flow on simplex (Byrne and Girolami, 2013) and coordinate system.

	HMC based on CORE			Geodesic HMC	Coord System HMC
	Approx $\lambda = 10^{-3}$	Approx $\lambda = 10^{-4}$	DA		
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 3: Average effective sample size per 1000 iterations in Dir(α), under different α.

6 Application: Finding Sparse Basis in a Population of Networks

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, recorded as an 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 between 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 for self-connectivity.

One scientific interest in neuroscience is to quantify the variation of brain networks and identify the regions (nodes) that contribute to it. Extending factor analysis to multiple matrices, one appealing approach is to have the networks share a common factor matrix but let the loadings vary across subjects. This can be considered as a simplified equivalent of three-way tensor factorization (Kolda and Bader, 2009). Then to selectively identify the important nodes, one natural way is to apply shrinkage on the elements of factor matrix.

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

To apply shrinkage in the constrained space, we adopt the induced prior as common in Bayesian literature (reviewed by Polson and Scott (2012)), which usually takes the form hierarchical structure $\theta_i \mid \kappa_i, \sigma \sim \text{No}(0, \kappa_i\sigma), \quad \kappa_i \sim G_1, \quad \sigma \sim G_2$ with κ_i, σ as the local and global scale parameters. However, when constraining θ_i, one caveat would be only adapting the conditional density $\text{No}(\theta_i; \kappa_i\sigma)$, which yields intractable normalizing constant involving $\kappa_i\sigma$ in the conditional. This difficulty can be avoided by reparameterizing $\theta_i = \eta_i\kappa_i\sigma$ with $\eta_i \sim \text{No}(0,1)$, and adapting the *joint* density of $\{\eta_i, \kappa_i, \sigma\}$ on constrained space instead. The joint density will not have intractable constant as long as the hyper-parameters in G_1 and G_2 are fixed.

We now take the Dirichlet-Laplace prior (Bhattacharya et al., 2015) as unconstrained distribution $\pi_{\mathcal{R}}$ and adapt it onto Stiefel manifold via (**??**).

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

$$\psi_{(i,k,l)} = \sum_{r=1}^{d} v_{(i,r)} u_{(k,r)} u_{(l,r)}$$

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

$$u_{(k,r)} = \eta_{(k,r)}\kappa_{(k,r)}\sigma_u$$

$$\eta_{(k,r)} \sim \text{Lap}(0,1), \quad \{\kappa_{(1,r)} \ldots \kappa_{(V,r)}\} \sim \text{Dir}(\alpha), \quad \sigma_u^2 \sim \text{IG}(2,1)$$

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

$$v_{(i,r)} \sim \text{No}(0, \sigma_{v,(r)}^2), \quad \sigma_{v,(r)}^2 \sim \text{IG}(2,1)$$

for $k > l$, $k = 2, \ldots, V$, $i = 1, \ldots, n$; Lap$(0, 1)$ denotes the Laplace distribution centered at 0 with scale 1; $Z = \{z_{(k,l)}\}_{k=1,\ldots,V;l=1,\ldots,V}$ is a symmetric unstructured matrix that serves as the latent mean; $\{v_{(i,r)}\}_{r=1,\ldots,d}$ is the loading for the ith network, with each $v_{(i,r)} > 0$; for all other scale parameters $\sigma_\cdot^2$, we choose weakly informative prior inverse Gamma IG$(2, 1)$, as appropriate for the scale under the logistic link. To induce sparsity in each Dirichlet, we use $\alpha = 0.1$ as suggested by Bhattacharya et al. (2015).

There are two types of constraints in the model, $U'U = I_d$ and $\sum_{k=1}^V \kappa_{(k,r)} = 1$ for $r = 1, \ldots, d$. Taking $v_1(U) = U'U - I_d$ and $v_2(\kappa_{(k,r)}) = \sum_{k=1}^V \kappa_{(k,r)} - 1$ for each r, the Jacobian is constant in (**??**). For posterior computation, we use DA-CORE as described above. Using latent variable w_U d-by-d upper triangular and positive diagonal matrix, and $w_{\kappa,(r)} > 0$ for $r = 1, \ldots, d$, we relax the parameters to

$$U^* = U w_U, \quad \kappa^*_{(k,r)} = \kappa_{(k,r)} w_{\kappa,(r)},$$

which yields re-parameterization via projection

$$\begin{aligned} &U = U^* w_U^{-1}, \quad w_U = \text{QR.R}(U^*), \\ &\kappa_{(k,r)} = \frac{\kappa^*_{(k,r)}}{w_{\kappa,(r)}}, \quad w_{\kappa,(r)} = \sum_{k=1}^V \kappa^*_{(k,r)} \\ &\eta_{k,r} = \frac{u_{(k,r)}}{\kappa_{(k,r)}\sigma_u}, \end{aligned} \tag{17}$$

where QR.R denotes the function that outputs R matrix in QR decomposition. To control the amount of relaxation, we assign w_U near I_d via $\pi(w_U) \propto \text{etr}\left[-\frac{(w_U - I_d)'(w_U - I_d)}{\lambda}\right]$ and $w_{\kappa,(r)}$ near 1 via $\pi(w_{\kappa,(r)}) \propto \exp\left[-\frac{(w_{\kappa,(r)} - 1)^2}{\lambda}\right]$ and set $\lambda = 10^{-3}$.

For comparison, we test with the specified model (i) against (ii) the same model except with simple $u_{(k,r)} \sim \text{No}(0, \sigma_u^2)$ instead of the shrinkage prior and (iii) the same model except without the orthonormality constraint $U'U = I$ and the shrinkage prior. We run all models for $10,000$ iterations and discard the first $5,000$ iteration as burn-in. For each iteration, we run 300 leap-frog steps. For efficient computing, we truncated $d = 20$.

Table 4 lists the benchmark results. Compared to (i) and (ii), the unconstrained model (iii) suffers from very low effective sample size, due to the serious convergence issue in the factor matrix U. As explained by previous findings in matrix/tensor factorization (Hoff et al., 2016), the factor matrix could scale and rotate without changing the likelihood, and substantial improvement could be obtained by applying orthonormality constraint.

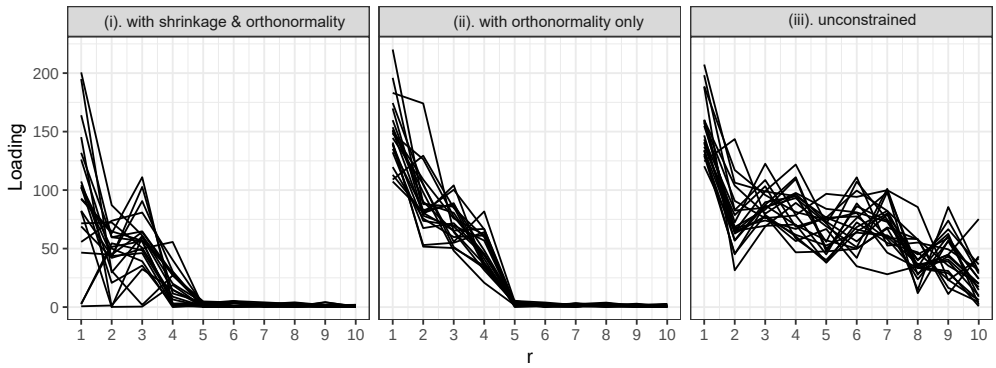
Figure 5(a) plots the posterior mean loadings $v_{(i,r)}$, with each line representing one subject. For all

$i = 1, \ldots, 21$, the lines drop quickly to near 0 after $r \geq 5$ in model (i) and (ii), but only do so until $r \geq 10$ in model (iii). This indicates that independent factors are more effective representation of the span, compared to non-orthogonal ones. Clearly, (i) shows more variability than (ii) in the loading $v_{(i,r)}$. We validate these models by calculating area under the receiver operating characteristic curve (AUC) based on the mean predicted probability and the binary outcome $A_{(i,k,l)}$, using the fitted data and the other unused rescan data from the 21 subjects. The models (i) and (ii) with orthonomality constraint perform similarly well, and clearly better than the unconstrained model (iii) in prediction AUC.

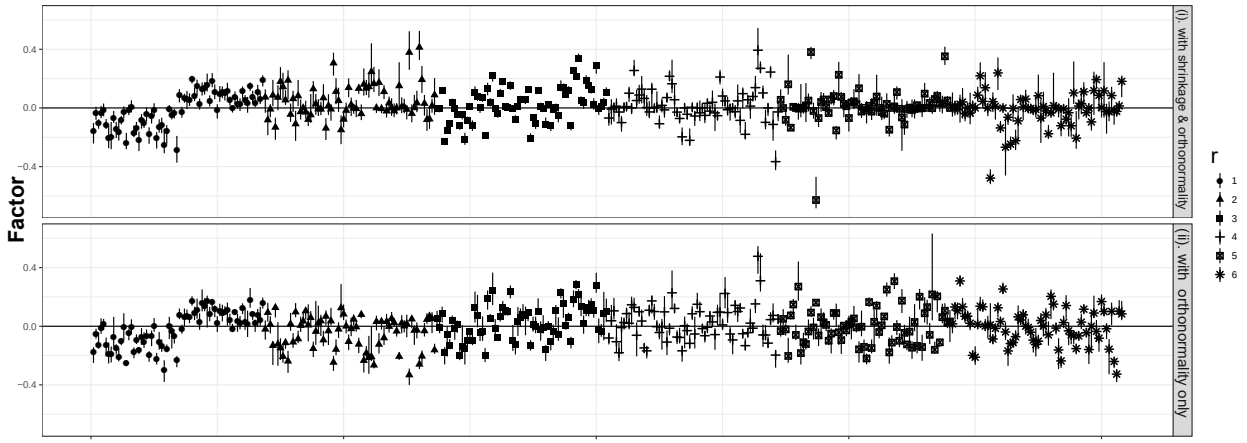
Model	(i).with shrinkage & orthonormality	(ii).with orthonormality only	(iii).unconstrained
Fitted AUC	97.9%	97.1%	96.9%
Prediction AUC	96.2%	96.2%	93.6%
ESS /1000 Iterations	193.72	188.10	8.15

Table 4: Comparing 3 models for 21 brain networks

Figure 5(b) compares the models (i) and (ii) over the top 6 frames of U_r, with r re-ordered such that $\sigma^2_{v,(1)} \geq \sigma^2_{v,(2)} \geq \ldots \geq \sigma^2_{v,(d)}$. The posterior of U_1, U_2, U_3 look very similar between the two, whereas U_4, U_5, U_6 have a considerable subset of points close to 0 in the model with shrinkage prior.



(a) Posterior mean of the loadings $v_{i,r}$ for 21 subjects using three models. Each line represents the loadings for one subject over $r = 1, \ldots 10$.



(b) Posterior mean and pointwise 95% credible interval of the factors $U_1, \ldots, U_6$ in the two constrained models.

Figure 5: Loadings and factors estimates of the network models. Panel (a) compares the varying loadings of the subjects in three models; Panel (b) compares the estimated shared factors with and without the shrinkage prior (model (iii) is omitted due to non-convergence in the factors).

7 Discussion

Parameter constraint often limits the flexibility to develop new model and creates huge burden in developing efficient posterior sampling algorithms. In this article, we develop a formal strategy to utilize the large pool of distributions in the constrained space, and propose a constraint relaxation approach to allow simple implementation for posterior estimation. For common constrained space that can be projected to via a function, we propose an exact algorithm based on data augmentation; for more general problem, we propose an approximation approach. This strategy works well for general equality and inequality constraints.

The future work of this research may include tackling the 'doubly intractable' problem. This issue is common when the data is on the constrained space, or the constrained prior has hyper-parameters to estimate. In the data application, we show that a reparameterization strategy works for some shrinkage priors, but clearly, more general treatment is needed. We expect our work to be compatible to the existing solutions (Murray et al., 2012; Rao et al., 2016; Stoehr et al., 2017).

A Proofs for Section 3.1

Proof. Proof of Lemma 1

Recall, that the distance function $v_d(\theta)$ is chosen so that $v_d(\theta)$ is zero for all $\theta \in \mathcal{D}$. It follows that for any function g

$$\begin{aligned}&\int_{\mathcal{R}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)\\&=\int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)+\int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta).\end{aligned} \tag{18}$$

Then,

$$\begin{aligned}&\left|E[g(\theta)|\theta\in\mathcal{D}]-E_{\tilde{\pi}_\lambda}[g(\theta)]\right|\\&=\left|\frac{\int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)}{\int_{\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)}-\frac{\int_{\mathcal{R}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp\big(-v_d(\theta)/\lambda\big)d\mu_{\mathcal{R}}(\theta)}{\int_{\mathcal{R}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp\big(-v_d(\theta)/\lambda\big)d\mu_{\mathcal{R}}(\theta)}\right|\\&=\left|\frac{\int_{\mathcal{R}\setminus\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)\cdot\int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)-\int_{\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\cdot\int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)}{\int_{\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)[\int_{\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)+\int_{\mathcal{R}\setminus\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(}\right|\end{aligned}$$

where the second equality follows from combining the fractions and making use of (3). We can bound the denominator from below by $\big[\int_{\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\big]^2>0$ so that

$$\begin{aligned}&\big|E[g(\theta)|\theta\in\mathcal{D}]-E_{\tilde{\pi}_\lambda}[g(\theta)]\big|\\&\leq\frac{\big|\int_{\mathcal{R}\setminus\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)\cdot\int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)-\int_{\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\cdot\int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(}{C_{\mathcal{D}}^2}\end{aligned}$$

where $C_{\mathcal{D}}=\int_{\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)$. If we add and subtract

$$\int_{\mathcal{R}\setminus\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)\cdot\int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)$$

within the numerator, we can apply the triangle inequality. Thus,

$$\begin{aligned}&\big|E[g(\theta)|\theta\in\mathcal{D}]-E_{\tilde{\pi}_\lambda}[g(\theta)]\big|\\&\leq\frac{\left|\int_{\mathcal{R}\setminus\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)\right|\cdot\left|\int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)-\int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\right|}{C_{\mathcal{D}}^2}\\&\quad+\frac{\left|\int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)\right|\cdot\left|\int_{\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)-\int_{\mathcal{R}\setminus\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu\right|}{C_{\mathcal{D}}^2}\end{aligned}$$

Since $g \in \mathbb{L}^1(\mathcal{R}, \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}d\mu_{\mathcal{R}})$, we can then bound the numerators as follows. First,

$$\begin{aligned}
&\left|\int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)\right| \cdot \left|\int_{\mathcal{D}} g(y_i)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) - \int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}\right| \\
&\leq \left|\int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)\right| \cdot \left(\left|\int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right| + \left|\int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\right.\right. \\
&\leq \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta) \cdot \left(\int_{\mathcal{D}} |g(y_i)|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) + \int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda\right. \\
&\leq \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta) \cdot \int_{\mathcal{R}} |g(x_i)|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) = C_{\mathcal{R}}E|g(x_i)| \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d
\end{aligned}$$

Here, $C_{\mathcal{R}} = \int_{\mathcal{R}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)$ is the normalizing constant of $\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}$. Secondly,

$$\begin{aligned}
&\left|\int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)\right| \cdot \left|\int_{\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) - \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)\right| \\
&\leq \int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)dx \cdot \left|\int_{\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right| + \left|\int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)\right| \\
&\leq \int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)\left(\int_{\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) + \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right) \\
&= \int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta).
\end{aligned}$$

Thus, we have the bounds specified by the theorem,

$$\begin{aligned}
&\left|E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)]\right| \\
&\leq \frac{C_{\mathcal{R}}E|g(\theta)| \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)}{C_{\mathcal{D}}^2} + \frac{\int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)}{C_{\mathcal{D}}^2} \\
&= \frac{\int_{\mathcal{R}\setminus\mathcal{D}} (C_{\mathcal{R}}E|g(\theta)| + |g(\theta)|)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)}{C_{\mathcal{D}}^2}.
\end{aligned}$$

It remains to be shown that

$$\left|E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)]\right| \to 0 \text{ as } \lambda \to 0^+.$$

Again, by the assumptions that $g \in \mathbb{L}^1(\mathcal{R}, \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}d\mu_{\mathcal{R}})$ and $v_d(\theta) > 0$ for $\mu_{\mathcal{R}}$ a.e. $\theta \in \mathcal{R}\setminus\mathcal{D}$, it follows that $(C_{\mathcal{R}}E|g(x_i)|+|g(x_i)|)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)$ is a dominating function of $(C_{\mathcal{R}}E|g(x_i)|+|g(x_i)|)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)$ which converges to zero for $\mu_{\mathcal{R}}$-a.e. $\theta \in \mathcal{R} \setminus \mathcal{D}$ as $\lambda \to 0^+$. Thus, by the dominated convergence theorem, $\left|E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)]\right| \to 0$ as $\lambda \to 0^+$.

□

Proof. Proof of Theorem 1

We begin with the bound from Lemma 1.

$$\left|E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)]\right| \leq \frac{\int_{\mathcal{R}\setminus\mathcal{D}}(C_{\mathcal{R}}E|g(\theta)| + |g(\theta)|)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)}{C_{\mathcal{D}}^2}.$$

For the moment, let us focus on the numerator of the previous expression. By the Cauchy-Schwartz inequality,

$$\begin{aligned}
&\int_{\mathcal{R}\setminus\mathcal{D}}(C_{\mathcal{R}}E|g(\theta)| + |g(\theta)|)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta) \\
&\leq \left(\int_{\mathcal{R}\setminus\mathcal{D}}(C_{\mathcal{R}}E|g(\theta)| + |g(\theta)|)^2\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right)^{1/2}\left(\int_{\mathcal{R}\setminus\mathcal{D}}\exp(-2v_d(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right)^{1/2} \\
&\leq \left(\int_{\mathcal{R}}(C_{\mathcal{R}}E|g(\theta)| + |g(\theta)|)^2\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right)^{1/2}\left(\int_{\mathcal{R}\setminus\mathcal{D}}\exp(-2v_d(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right)^{1/2}
\end{aligned}$$

By assumption, $g \in \mathbb{L}^2(\mathcal{R}, \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}\mu_{\mathcal{R}})$. Thus,

$$\begin{aligned}
&\int_{\mathcal{R}\setminus\mathcal{D}}(C_{\mathcal{R}}E|g(\theta)| + |g(\theta)|)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_d(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta) \\
&= \underbrace{\left([C_{\mathcal{R}}^3 + 2C_{\mathcal{R}}^2](E|g|)^2 + C_{\mathcal{R}}E[|g|^2]\right)^{1/2}}_{C_{\mathcal{R},g}<\infty}\left(\int_{\mathcal{R}\setminus\mathcal{D}}\exp(-2v_d(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right)^{1/2} \\
&= C_{\mathcal{R},g}\left(\int_{\mathcal{R}\setminus\mathcal{D}}\exp(-2v_d(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right)^{1/2}
\end{aligned}$$

We separate the integral

$$\int_{\mathcal{R}\setminus\mathcal{D}}\exp(-2v_d(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)$$

over the sets $\{\theta : v_d(\theta) > -\lambda\log\lambda\}$ and $\{\theta : 0 < v_d(\theta) < -\lambda\log\lambda\}$.

$$\begin{aligned}
&\int_{\mathcal{R}\setminus\mathcal{D}}\exp(-2v_d(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) \\
&= \int_{\{\theta:v_d(\theta)>-\lambda\log\lambda\}}\exp(-2v_d(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) + \int_{\{\theta:\,0<v_d(\theta)<-\lambda\log\lambda\}}\exp(-2v_d(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) \\
&\leq \lambda^2\int_{\{\theta:v_d(\theta)>-\lambda\log\lambda\}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) + \int_{\{\theta:\,0<v_d(\theta)<-\lambda\log\lambda\}}\exp(-2v_d(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) \\
&\leq C_{\mathcal{R}}\lambda^2 + \int_{\{\theta:\,0<v_d(\theta)<-\lambda\log\lambda\}}\exp(-2v_d(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)
\end{aligned}$$

To review, to this point we have shown that

$$\left|E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)]\right| \leq \frac{C_{\mathcal{R},g}}{D_{\mathcal{D}}^2}\left(C_{\mathcal{R}}\lambda^2 + \int_{\{\theta:\, 0<v_d(\theta)<-\lambda\log\lambda\}} \exp(-2v_d(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right)^{1/2} \tag{19}$$

From the requirements of Theorem 1, we now let $v_d(\theta) = \inf_{x\in\mathcal{D}} ||\theta - x||_2$ and assume that $\mathcal{D}$ has a piecewise smooth boundary. In this case, the set $\{\theta : 0 < v_d(\theta) < -\lambda\log\lambda\}$ forms a 'shell' of thickness $-\lambda\log\lambda$ which encases $\mathcal{D}$.

For the moment, suppose that $\mathcal{D}$ is a bounded subset of $\mathcal{R}$. Furthermore, suppose we take λ sufficiently small so that $\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)$ is continuous on $V_\lambda = \{\theta : 0 < v_d(\theta) < -\lambda\log\lambda\}$. Observe that

$$\int_{\{\theta:\, 0<v_d(\theta)<-\lambda\log\lambda\}} \exp(-2v_d(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) \leq \sup_{V_\lambda}|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)|\int_{V\lambda}\exp(-2v_d(\theta)/\lambda)d\mu_R(\theta)$$

$$\leq \sup_{V_\lambda}|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)|\int_{V\lambda} d\mu_R(\theta) = \sup_{V_\lambda}|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)|\cdot Vol(V_\lambda)$$

$$= \sup_{V_\lambda}|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)|S_{\mathcal{D}}\cdot\lambda|\log\lambda|$$

Here,$S_{\mathcal{D}}$ is the surface area of boundary of $\mathcal{D}$, which is finite by the assumptions that $\mathcal{D}$ is bounded and has a piecewise smooth boundary. Additionally, since V_λ is relatively compact, it follows that $\sup_{V_\lambda}|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)| < \infty$.

Consider the more general case where $\mathcal{D}$ is not a bounded subset of $\mathcal{R}$. Since $\int_{\mathcal{R}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)$, there exists a radius ρ such that $\int_{||\theta||_2>\rho}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) < \lambda^2$. Note that, for $\theta \in V_\lambda$, $J(v_d(\theta)) = \sqrt{(Dv_d)'(Dv_d)} = 2$. By the co-area formula Diaconis et al. (2013); Federer (2014)

$$\int_{\{\theta:\, 0<v_d(\theta)<-\lambda\log\lambda\}} \exp(-2v_d(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) = \int_0^{-\lambda\log\lambda} e^{-\frac{x}{\lambda}}\left(\int_{v_d^{-1}(x)}\frac{1}{2}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta)\right)dx$$

Again, we may take λ sufficiently small so that $\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)$ is continuous on V_λ. As such, the function $\int_{v_d^{-1}(x)}\frac{1}{2}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta)$ is a continuous map from the closed interval, $[0, -\lambda\log\lambda]$, to $\mathbb{R}$. Hence it is bounded. As a result,

$$\int_{\{\theta:\, 0<v_d(\theta)<-\lambda\log\lambda\}} \exp(-2v_d(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)$$

$$\leq \sup_{x\in[0,-\lambda\log\lambda]}\left(\int_{v_d^{-1}(x)}\frac{1}{2}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta)\right)\cdot\int_0^{-\lambda\log\lambda} e^{-\frac{x}{\lambda}}dx$$

$$= \sup_{x\in[0,-\lambda\log\lambda]}\left(\int_{v_d^{-1}(x)}\frac{1}{2}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta)\right)\cdot(\lambda-\lambda^2) = O(\lambda)$$

This result also applies to the case where $\mathcal{D}$ is bounded. Thus, we may conclude that

$$\begin{aligned}
&\left|E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)]\right| \\
&\leq \frac{C_{\mathcal{R},g}}{D_{\mathcal{D}}^2}\left(C_{\mathcal{R}}\lambda^2 + \sup_{x\in[0,-\lambda\log\lambda]}\left(\int_{v_d^{-1}(x)}\frac{1}{2}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta)\right)\cdot(\lambda-\lambda^2)\right)^{1/2} \\
&= \frac{C_{\mathcal{R},g}}{D_{\mathcal{D}}^2}\cdot\sup_{x\in[0,-\lambda\log\lambda]}\left(\int_{v_d^{-1}(x)}\frac{1}{2}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta)\right)\cdot\sqrt{\lambda}+o(\sqrt{\lambda})
\end{aligned}$$

Since $\sup_{x\in[0,-\lambda\log\lambda]}\left(\int_{v_d^{-1}(x)}\frac{1}{2}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta)\right)$ is a decreasing function in λ, we may conclude that $\left|E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)]\right| = O(\sqrt{\lambda})$. □

B Proofs from Section 3.2

Proof. Recall that we have two densities. The first is the fully constrained density for $\theta \in \mathcal{D}$.

$$\pi_{\mathcal{D}}(\theta) = \frac{1}{m_0}\frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))}\mathbb{1}_{\mathcal{D}}(\theta)$$

where the normalizing constant m_0 is calculated w.r.t. Hausdorff measure

$$m_0 = \int_{\mathcal{R}}\frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))}\mathbb{1}_{\mathcal{D}}(\theta)d\bar{\mathcal{H}}^{r-s}(\theta).$$

Secondly, we have the relaxed distribution

$$\tilde{\pi}_{\mathcal{D}}(\theta) = \frac{1}{m_\lambda}\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)\exp\left(-\frac{||v(\theta)||_1}{\lambda}\right)$$

where the normalizing constant is calculated w.r.t. Lebesgue measure on $\mathcal{R}$, denote by $\mu_{\mathcal{R}}$,

$$m_\lambda = \int_{\mathcal{R}}\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)\exp\left(-\frac{||v(\theta)||_1}{\lambda}\right)d\mu_{\mathcal{R}}(\theta).$$

For a given function, $g : \mathcal{R} \to \mathbb{R}$, we can define the exact and approximate expectations of g, respectively

E_Π and $E_{\tilde{\Pi}}$, as

$$E_\Pi[g(\theta)] = E[g(\theta)|\theta \in \mathcal{D}] = \int_{\mathcal{R}} \frac{g(\theta)}{m_0} \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} \mathbb{1}_{\mathcal{D}}(\theta) d\bar{\mathcal{H}}^{r-s}(\theta)$$

$$= \int_{\mathcal{D}} \frac{g(\theta)}{m_0} \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta)$$

$$E_{\tilde{\Pi}}[g(\theta)] = \int_{\mathcal{R}} \frac{g(\theta)}{m_\lambda} \mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta) \exp\left(- \frac{||v(\theta)||_1}{\lambda} \right) d\mu_{\mathcal{R}}(\theta)$$

$$= \int_{\mathbb{R}^s} \frac{1}{m_\lambda} \int_{\nu^{-1}(x)} g(\theta) \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta)}{J(\nu(\theta))} \exp\left(- \frac{||\nu(\theta)||_1}{\lambda} \right) d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}$$

$$= \int_{\mathbb{R}^s} \frac{\exp\left(- \frac{||x||_1}{\lambda} \right)}{m_\lambda} \int_{\nu^{-1}(x)} g(\theta) \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}$$

Let,

$$m(x) = m^{r-s}(x) = \int_{\nu^{-1}(x)} \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta).$$

By construction, $m(x) > 0$ for $\mu_{\mathbb{R}^s}$-a.e. $x \in Range(\nu)$. In particular, $m_0 = m(0) > 0$. By Theorem 1,

$$E[g(\theta)|\nu(\theta) = x] = \frac{1}{m(x)} \int_{\nu^{-1}(x)} g(\theta) \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta). \tag{20}$$

As such, we may may express $E_{\tilde{\Pi}}[g(\theta)]$ as

$$E_{\tilde{\Pi}}[g(\theta)] = \int_{\mathbb{R}^s} \frac{m(x)}{m_\lambda} \exp\left(- \frac{||x||_1}{\lambda} \right) E\big[g(\theta)|\nu(\theta) = x\big] d\mu_{\mathbb{R}^s}(x). \tag{21}$$

Let us first consider the small λ behavior of m_λ. We begin by re-expressing m_λ in terms of $m(x)$ through the co-area formula.

$$m_\lambda = \int_{\mathcal{R}} \pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta) \exp\left(- \frac{||v(\theta)||_1}{\lambda} \right) d\mu_{\mathcal{R}}(\theta)$$

$$= \int_{\mathbb{R}^s} \exp\left(- \frac{||x||_1}{\lambda} \right) \int_{\nu^{-1}(x)} \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}(x)$$

$$= \int_{\mathbb{R}^s} m(x) \exp\left(- \frac{||x||_1}{\lambda} \right) d\mu_{\mathbb{R}^s}(x)$$

Split the above integral into two regions: the interior and exterior of $B_1(0; \lambda|\log(\lambda^{s+1})|)$. Note that

outside of B_1, $\exp(-||x||_1/\lambda) \leq \lambda^{s+1}$.

$$
\begin{aligned}
m_\lambda &= \int_{\mathbb{R}^s \setminus B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) \exp\bigg(- \frac{||x||_1}{\lambda}\bigg) d\mu_{\mathbb{R}^s}(x) + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) \exp\bigg(- \frac{||x||_1}{\lambda}\bigg) d\mu_{\mathbb{R}^s}(x) \\
&= O\bigg(\lambda^{s+1} \int_{\mathbb{R}^s \setminus B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) d\mu_{\mathbb{R}^s}(x)\bigg) \\
&\qquad + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) \bigg[1 + O\bigg(\frac{1}{\lambda} \exp\bigg(- \frac{||x||_1}{\lambda}\bigg)\bigg)\bigg] d\mu_{\mathbb{R}^s}(x) \\
&= O\bigg(\lambda^{s+1} \int_{\mathbb{R}^s \setminus B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) d\mu_{\mathbb{R}^s}(x)\bigg) \\
&\qquad + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) \bigg[1 + O(\lambda^s)\bigg] d\mu_{\mathbb{R}^s}(x)
\end{aligned}
$$

Since $m(x)$ is continuous on an open neighborhood containing the origin, we may choose λ small enough so that $m(x)$ is uniformly continuous on $B_1(0;\lambda|\log\lambda^{s+1}|)$. Then,

$$
\begin{aligned}
m_\lambda &= O\bigg(\lambda^{s+1}\bigg) + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} [m(0) + o(1)][1 + O(\lambda^s)] d\mu_{\mathbb{R}^s}(x) \\
&= O(\lambda^{s+1}) + [m(0) + o(1)][1 + O(\lambda^s)] \underbrace{\frac{|2(s+1)\lambda\log\lambda|^s}{\Gamma(s+1)}}_{Vol(B_1(0;\lambda|\log(\lambda^{s+1})|)} \\
&= m(0)\frac{|2(s+1)\lambda\log\lambda|^s}{\Gamma(s+1)} + o(|\lambda\log\lambda|^s)
\end{aligned}
$$

at leading order as $\lambda \to 0^+$.

We know turn to the small λ behavior of $\tilde{E}[g(\theta)]$. Again, we may choose λ sufficient small so that both

$$
m(x) \int_{\nu^{(-1)}(x)} \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta)
$$

$$
G(x) = \int_{\nu^{(-1)}(x)} g(\theta) \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) = m(x) E[g|\nu(\theta) = x]
$$

are continuous on $B_1(0;\lambda|\log\lambda^{s+1}|)$ and hence uniformly continuous at $x = 0$.

Similar to the study of m_λ, separate the $\tilde{E}[g(\theta)]$ into integrals over the interior and exterior of $B_1(0, \lambda|\log(\lambda)^{s+1}|)$. Again, we assume λ is taken to be sufficiently small so that both $m(x)$ and $G(x)$ are uniformly continuous

on B_1. Then

$$
\begin{aligned}
E_{\tilde{\Pi}}[g(\theta)] &= \int_{\mathbb{R}^s} \frac{m(x)}{m_\lambda} \exp\left(- \frac{||\nu(x)||_1}{\lambda} \right) E[g(\theta)|\nu(\theta) = x] d\mu_{\mathbb{R}^s}(x) \\
&= \int_{\mathbb{R}^s} \frac{1}{m_\lambda} \exp\left(- \frac{||\nu(x)||_1}{\lambda} \right) \int_{\nu^{(-1)}(x)} \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}(x) \\
&= \int_{\mathbb{R}^s \setminus B_1(0;\lambda|\log(\lambda^{s+1})|)} \frac{1}{m_\lambda} \exp\left(- \frac{||x||_1}{\lambda} \right) \int_{\nu^{(-1)}(x)} \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}(x) \\
&\quad + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} \frac{1}{m_\lambda} \exp\left(- \frac{||x||_1}{\lambda} \right) \int_{\nu^{(-1)}(x)} \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}(x) \\
&= O\left(\frac{\lambda^{s+1}}{m_\lambda} \right) + \int_{B_1} \frac{m(0) + o(1)}{m_\lambda} (1 + O(\lambda^s)] \Big(E[g(\theta)|\nu(\theta) = 0] + o(1) \Big) d\mu_{\mathbb{R}^s}(x) \\
&= O\left(CE|g| \frac{\lambda}{|\log \lambda|^s} \right) + E[g(\theta)|\theta \in \mathcal{D}] + o(1).
\end{aligned}
$$

And we may conclude that

$$
\left| E[g|\theta \in \mathcal{D} - E_{\tilde{\Pi}}[g] \right| \to O \text{ as } \lambda \to 0^+.
$$

The proof of the corollary follows from changing the $o(1)$ correction within the integrals over $B_1(0;\lambda|\log \lambda^{s+1}|)$ with $O(\lambda|\log \lambda^{s+1}|)$ corrections. □

References

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*.

Bhattacharya, A., D. Pati, N. S. Pillai, and D. B. Dunson (2015). Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association 110*(512), 1479–1490.

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.

Hankin, R. K. et al. (2010). A generalization of the dirichlet distribution. *Journal of Statistical Software 33*(11), 1–18.

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.

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association 96*(453), 161–173.

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.

Kolda, T. G. and B. W. Bader (2009). Tensor decompositions and applications. *SIAM review 51*(3), 455–500.

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.

Murray, I., Z. Ghahramani, and D. MacKay (2012). Mcmc for doubly-intractable distributions. *arXiv preprint arXiv:1206.6848*.

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*.

Polson, N. G. and J. G. Scott (2012). Local shrinkage rules, lévy processes and regularized regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(2), 287–311.

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*.

Wang, C. and D. M. Blei (2009). Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In *Advances in neural information processing systems*, pp. 1982–1989.

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